Supplementary Information

Figure Captions

SI Figure 1. Orthogonal NMF produces low-error reconstruction Orthogonal nonnegative matrix factorization factors each a) gene expression data matrix into b) F, a set of orthogonal linear feature vectors, and c) C, a matrix of coefficients. For (b) we order the genes according to their weight in each feature, keeping only genes with greater $> 4\sigma$ for each feature. This ordering highlights the orthogonality of features. (d) The reconstructed data FC is a denoised version of the original data matrix. (e) Data values from the original data matrix are partitioned into bins, and their means are plotted against corresponding data values from the reconstructed data matrix (FC). Bin width ~ 0.1. Shaded area indicates the standard deviation between original values and reconstructed values within each bin. (See Methods - oNMF error analysis). (f) Histogram of the entries in normalized data matrix D, with log(counts) on y-axis. 99.996% of data values are less than 5, which is also the linear regime of (e). (g) Zoomed in view of white inset from (c), showing specific genes that have high weights for two features, as well as their top gene set results from gene set enrichment analysis.

SI Figure 2. Choosing dimensionality of feature set Plot of loss function for different feature sets, built by sweeping over many values of m (x-axis). Feature set with lowest loss (red asterisk) is chosen. This example uses data from Fig. 4. Loss function is a function of the residual error and a penalty on m. For loss function equation, see Methods - Extraction of gene feature vectors with matrix factorization.

SI Figure 3. PopAlign models for all 12 tissues of Tabula Muris For each tissue, we plot the following: (upper left) joint t-SNE plot of experimental single-cell data (black), PopAlign model-generated data (teal), and mixture model centroids (μ) as numbered disks, (upper right) Heatmap of the proportion of cells classified by each mixture model component against their annotated labels. Each column sums to 1. (lower) Heatmap of mixture component centroids (mu) in terms of their expression level of features. For all tissues, we used a universal 30D feature set determined using sampled data from all tissues within the collection.

SI Figure 4. Mixture components across Tabula Muris tissues are highly specific for single cell type Histogram of each mixture component's highest score for an annotated label. These highest scores as taken as the max of each column in Fig 3c,d and analogous cell annotation heatmaps for all tissues in SI Figure 3. Most mixture components score for a single annotated label uniquely.

SI Figure 5. Pairwise alignments between PopAlign models of Mammary Gland and Limb Muscle can be dissected in terms of Δw, $\Delta\mu$, $\Delta\Sigma$ Aligned subpopulations between the reference tissue, Mammary Gland, and the test tissue, Limb Muscle, are ranked by their Jeffrey's Divergence (a). For each aligned pair, we display: (b) associated p-value, (c) mean gene expression states (μ_i) in terms of annotated features, (d) shifts in abundance (Δw), (e) shifts in mean gene expression state ($\Delta\mu$), (f) shifts in population spread ($\Delta\Sigma$).

SI Figure 6. Pairwise alignments between PopAlign models of immune cells and signal conditions (IFNG, GM-CSF) can be dissected in terms of Δw, $\Delta\mu$, $\Delta\Sigma$ Aligned subpopulations between the reference PBMC sample and each test sample, GM-CSF and IFNG, are ranked by their Jeffrey's Divergence (a). For each aligned pair, we display: (b) mean gene expression states (μ_i) in terms of annotated features, (c) shifts in abundance (Δw), (d) shifts in mean gene expression state ($\Delta\mu$), (e) shifts in population spread ($\Delta\Sigma$).

SI Figure 7. Classification of reference patient population (healthy1) using canonical immune system markers. We classify cells using heatmap of mean gene expression values for cells from each reference subpopulation from healthy1. Each gene is independently scaled by its max value across all subpopulation means. Erythrocytes are HBB+, T cells are CD3D+, naive T cells are CCR+/CD62L+/CD27+/CD28+, effector T cells are CD57+/PD1+/CD95+, B cells are CD19+, and monocytes are CD14+/CD16+.

SI Figure 8. Pairwise alignments between PopAlign models of healthy1 and healthy2 Aligned subpopulations between the reference healthy1 sample and test sample, healthy2, are ranked by their Jeffrey's Divergence. For each aligned pair, the following are displayed: mean gene expression states (μ_i) in terms of features, shifts in abundance (Δw), shifts in mean gene expression state ($\Delta\mu$), shifts in population spread ($\Delta\Sigma$), and log10(Jeffrey's

Divergence).

SI Figure 9. Pairwise alignments between PopAlign models of healthy1 and MM1 See caption for SI Figure 8.

SI Figure 10. Pairwise alignments between PopAlign models of healthy1 and MM2 See caption for SI Figure 8.

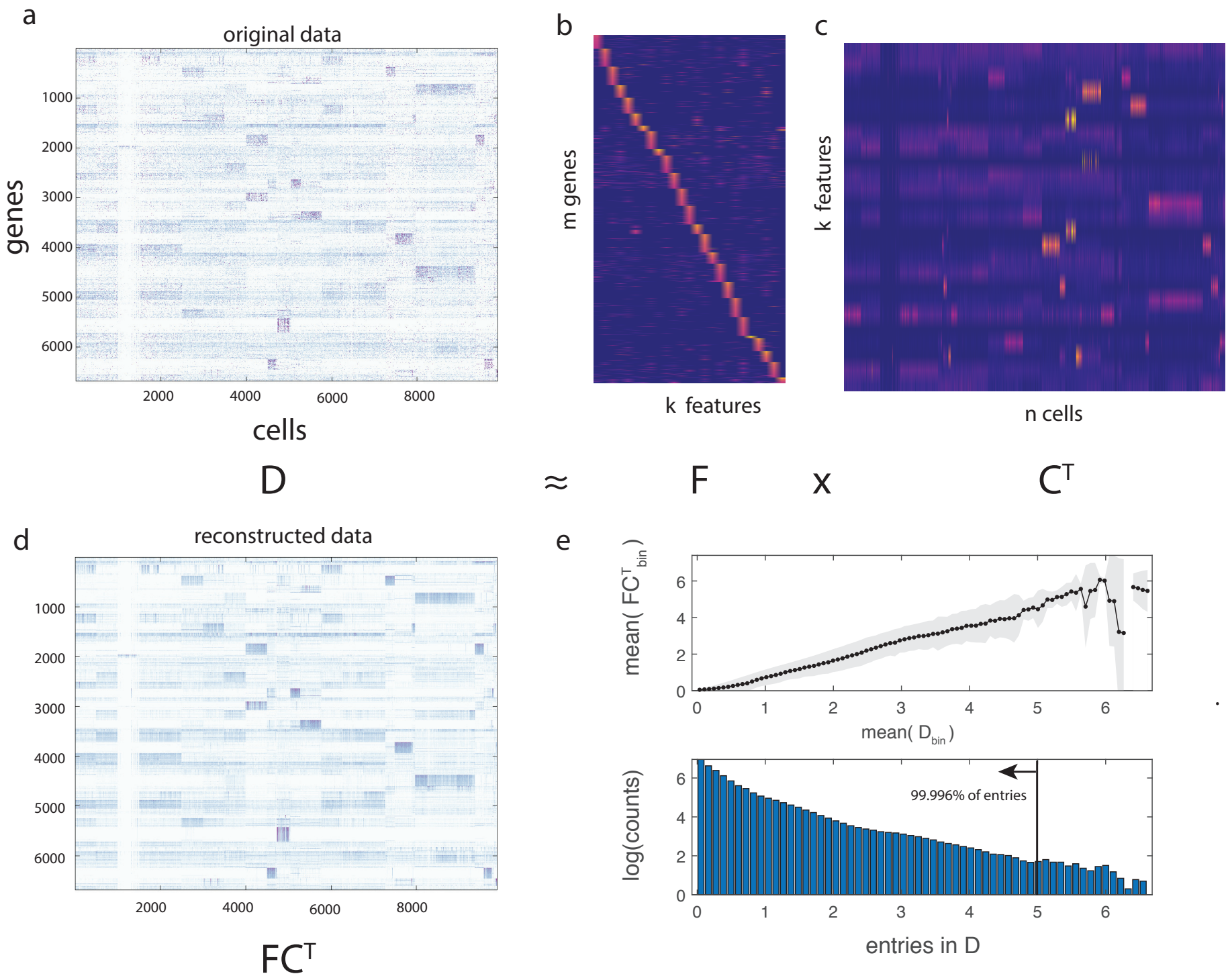
SI Figure 11. Pairwise alignments between PopAlign models of healthy1 and MM3 See caption for SI Figure 8.

SI Figure 12. Pairwise alignments between PopAlign models of healthy1 and MM4 See caption for SI Figure 8.

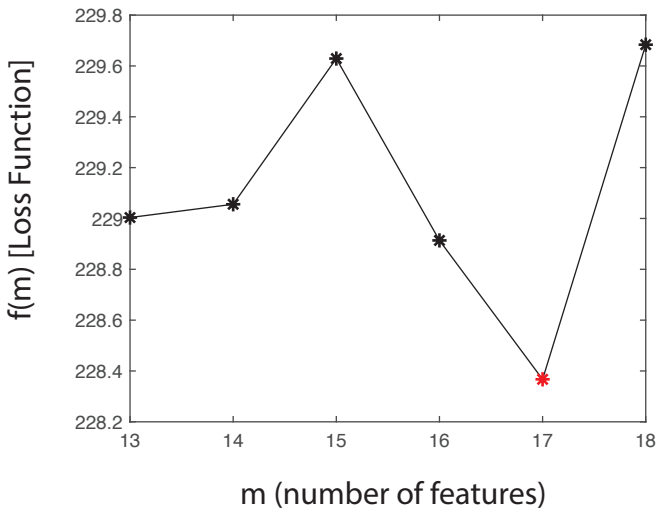
SI Figure 13. t-SNE analysis obscures the relationships between aligned subpopulations (a-c) tSNE plots of single cell data for three signaling conditions from Figure 4: (a) PBMC, (b) GM-CSF, and (c) IFNG. Mixture component $\mu's$ (disks) for PBMC mixture components 1 and 3 overlaid on plots (b) and (c) to aid interpretation. PopAlign generated alignments for monocyte mixtures from figure (3) are indicated as lines. tSNE was performed on data pooled from all experiments, but experimental conditions are shown individually for clarity. Plot shows that gene expression changes due to signaling can result in relative cluster shifts within the tSNE plots that obscure subpopulation alignment. For example, in (b) the monocyte populations aligned by PopAlign (Figure 4) become separated by a cluster of T-cells (mixture components 3 and 4).

SI Table 1. Patient information for PBMC donors Patient information table with Age, Gender, Ethnicity, Disease Status, and Medications.

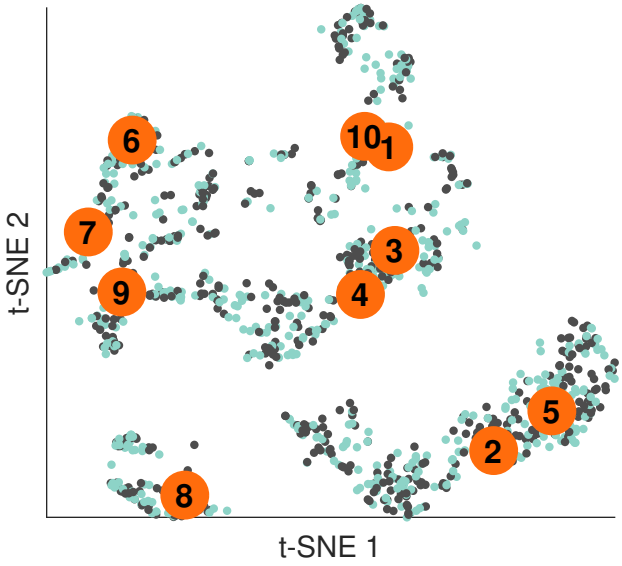
SI Figure 1

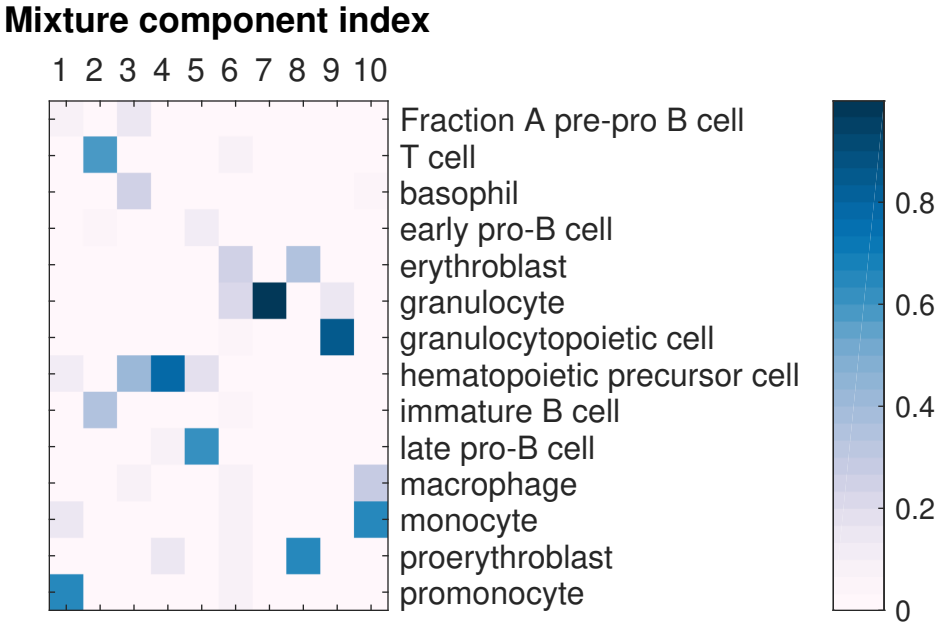

a
original data
genes
cells
1000
2000
3000
4000
5000
6000
2000
4000
6000
8000
b
m genes
k features
c
k features
n cells
D
≈
F
x
C^T
d
reconstructed data
1000
2000
3000
4000
5000
6000
2000
4000
6000
8000
FC^T
e
mean(FC^T_bin)
mean(D_bin)
log(counts)
99.996% of entries
entries in D

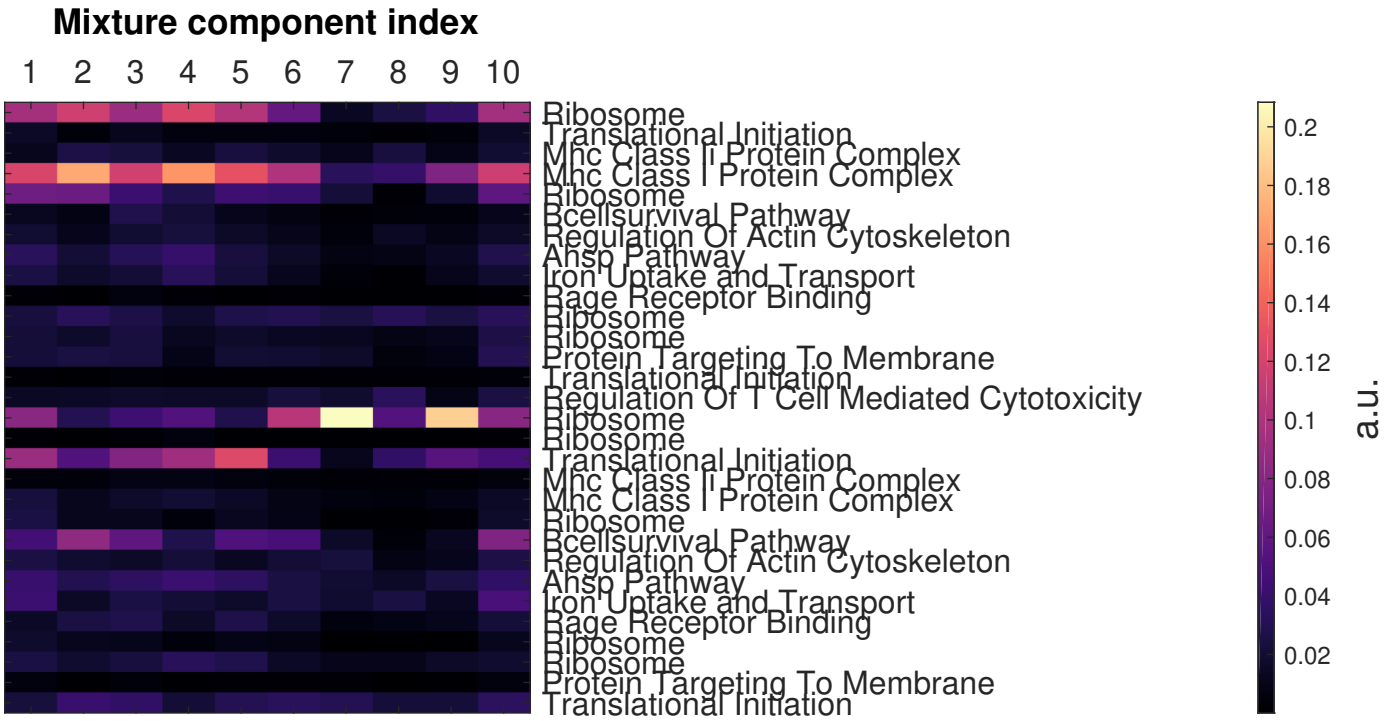
SI Figure 2


f(m) [Loss Function]
229.8
229.6
229.4
229.2
229
228.8
228.6
228.4
228.2
13
14
15
16
17
18
m (number of features)

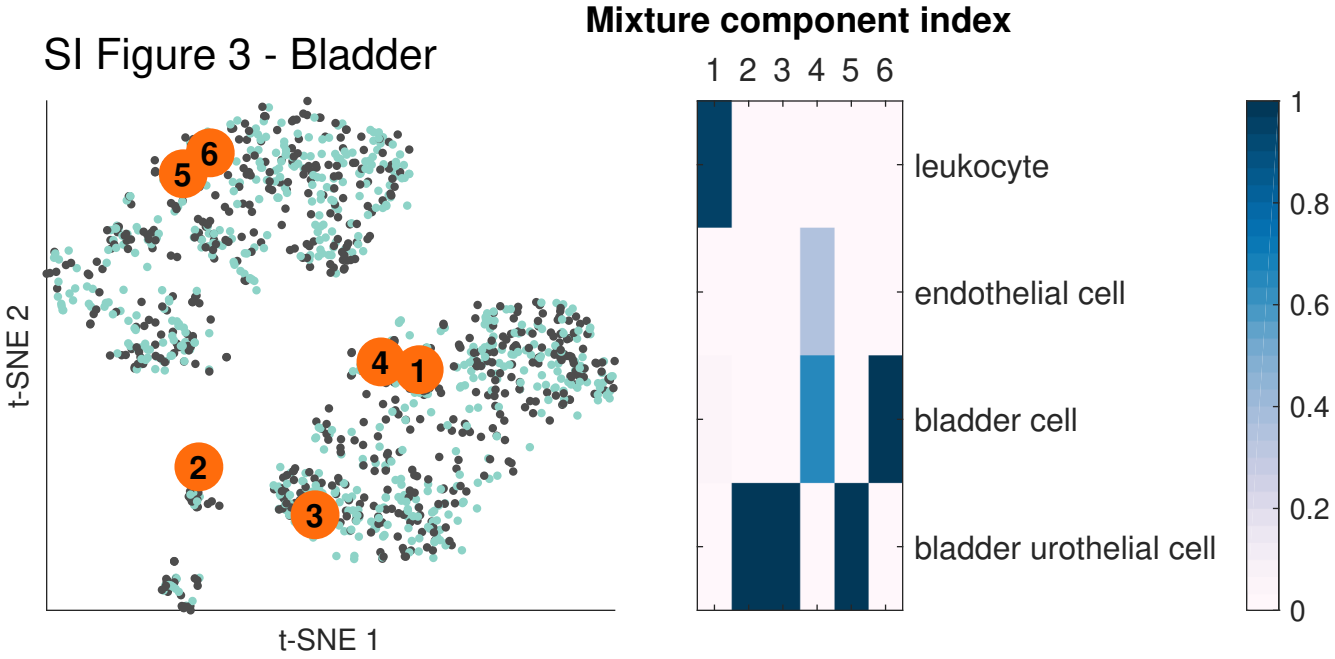
SI Figure 3 - Marrow

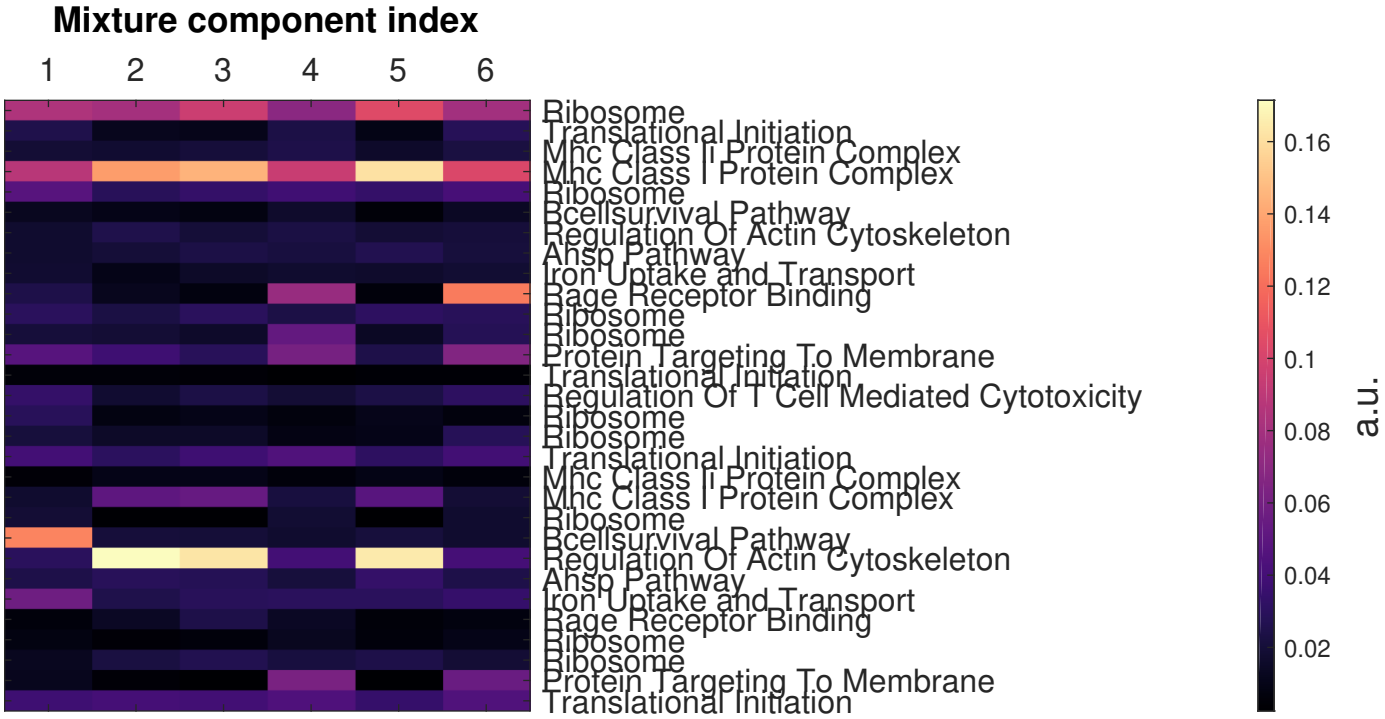




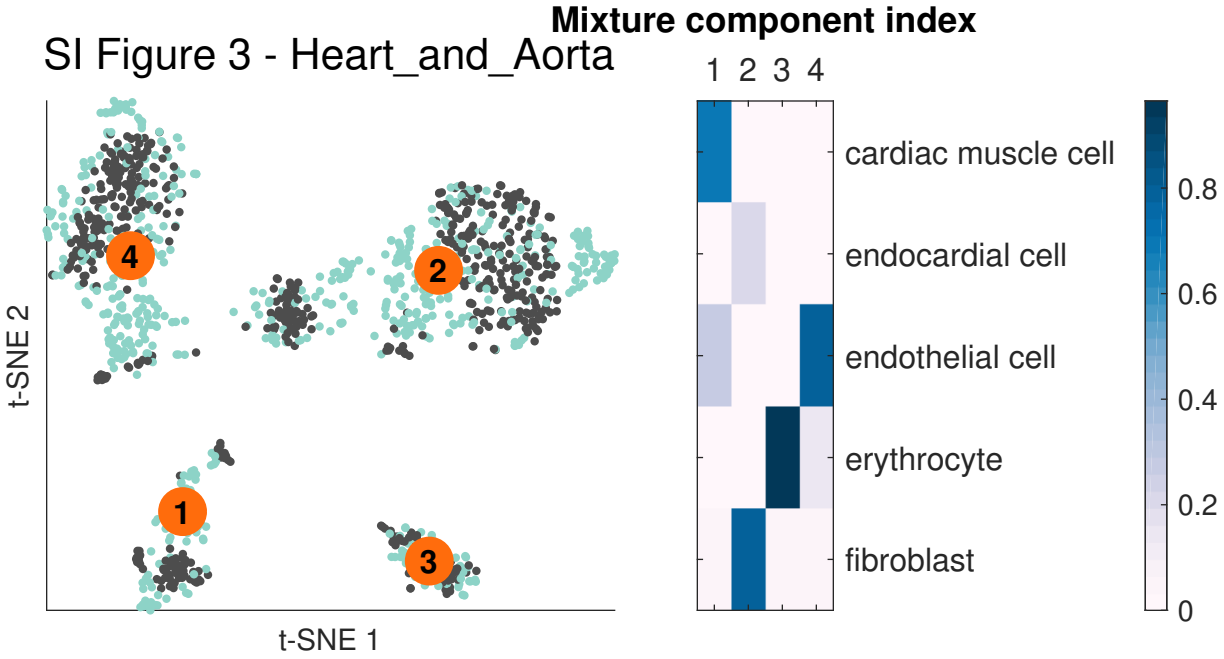


SI Figure 3 - Bladder

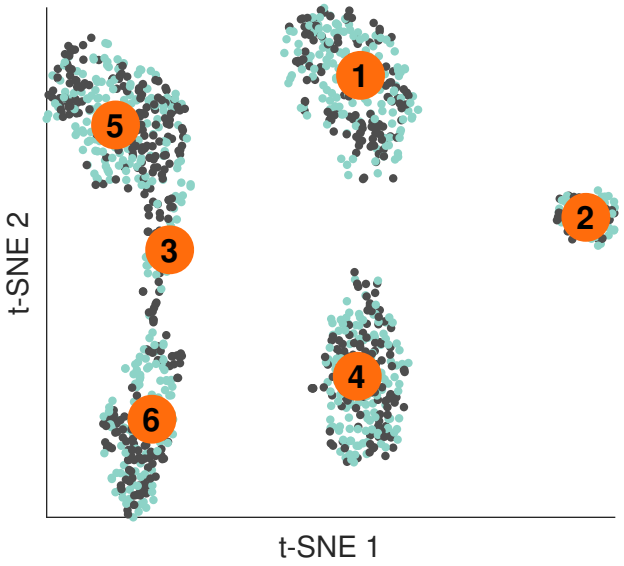


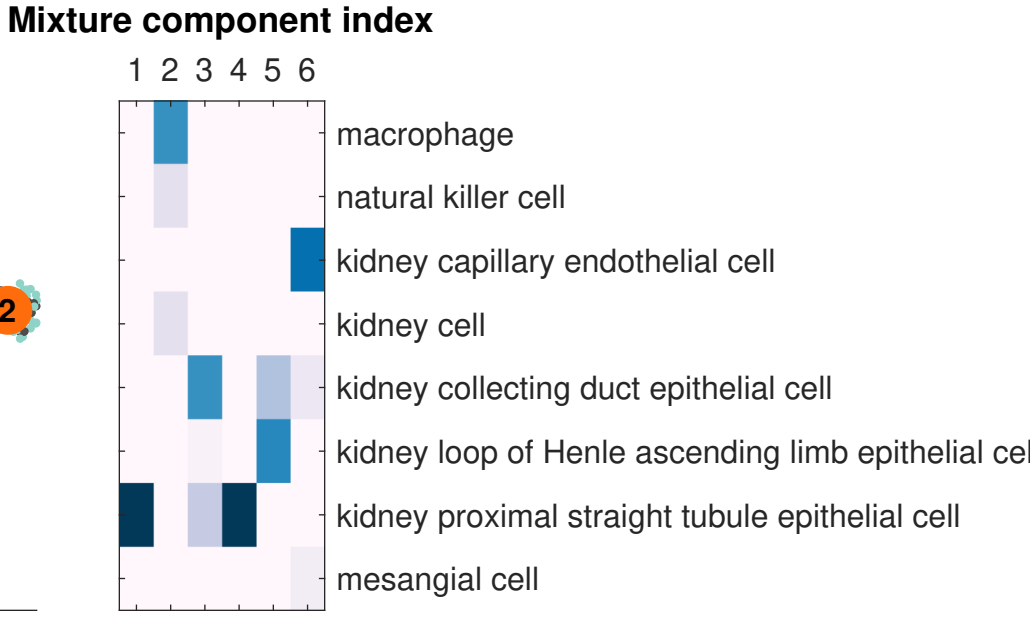


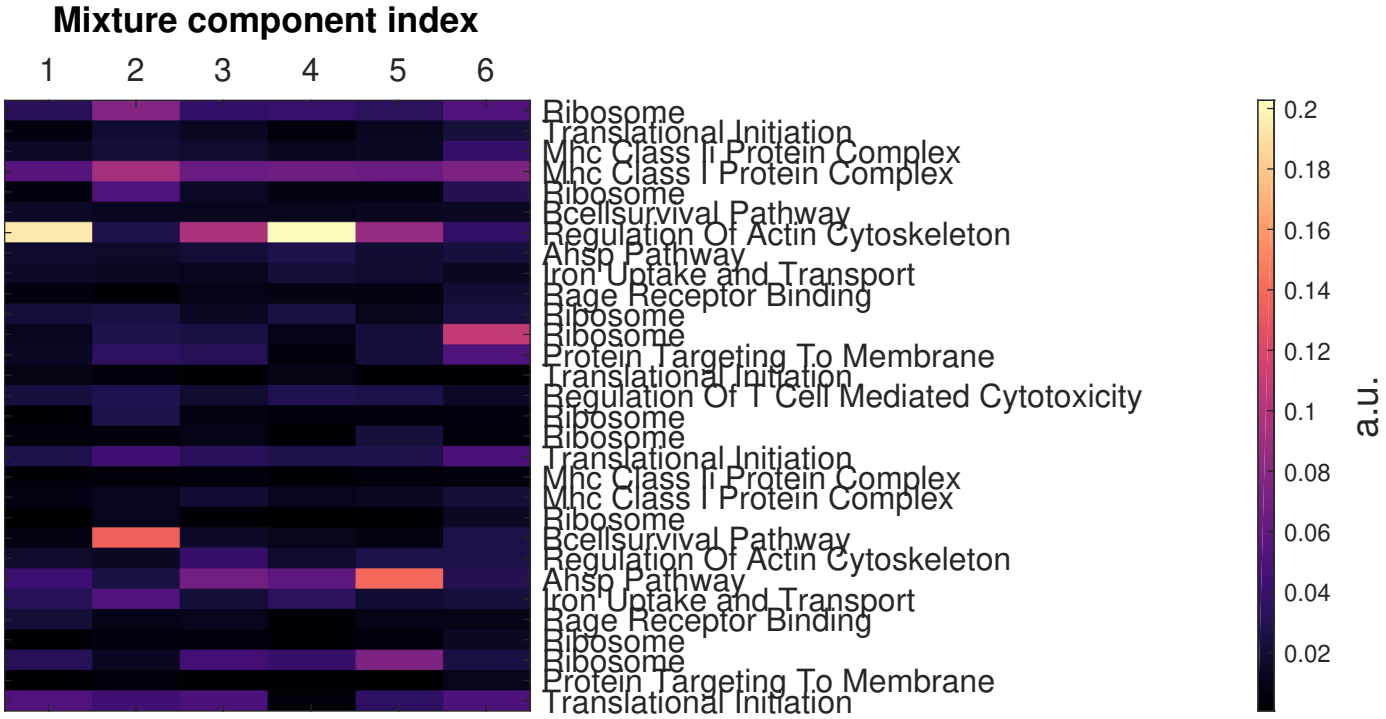
SI Figure 3 - Heart_and_Aorta



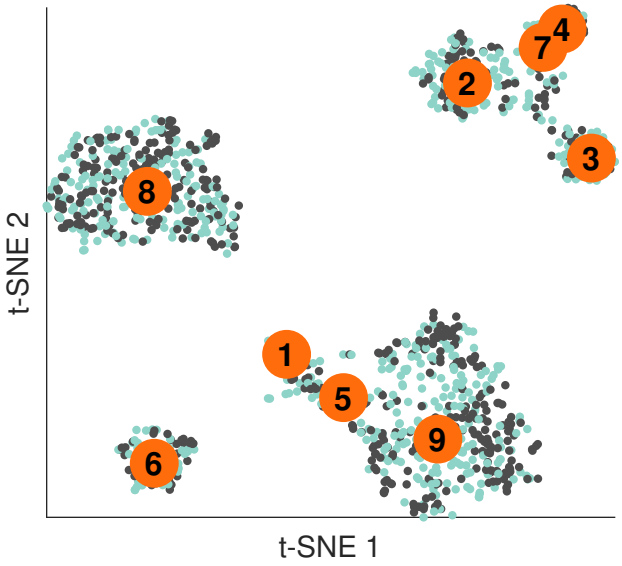
SI Figure 3 - Kidney

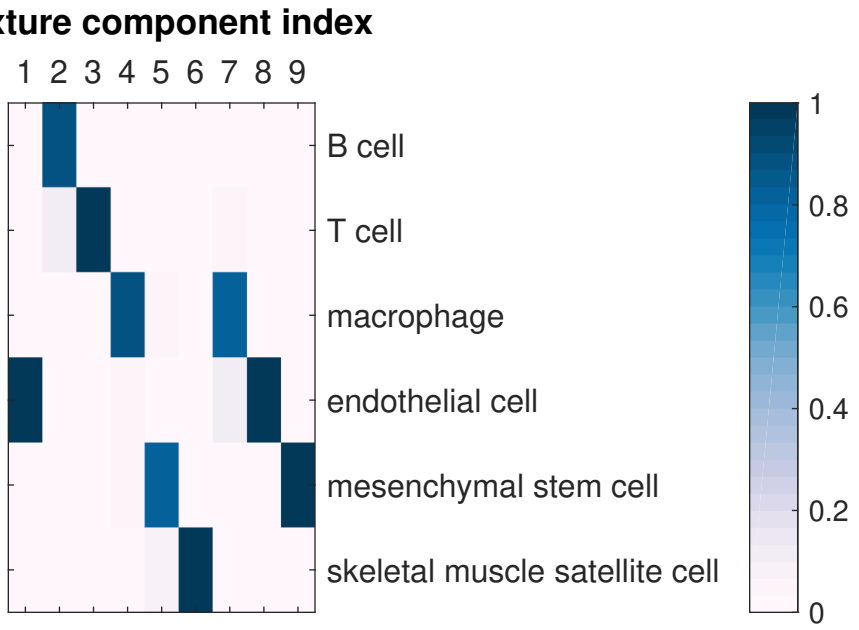


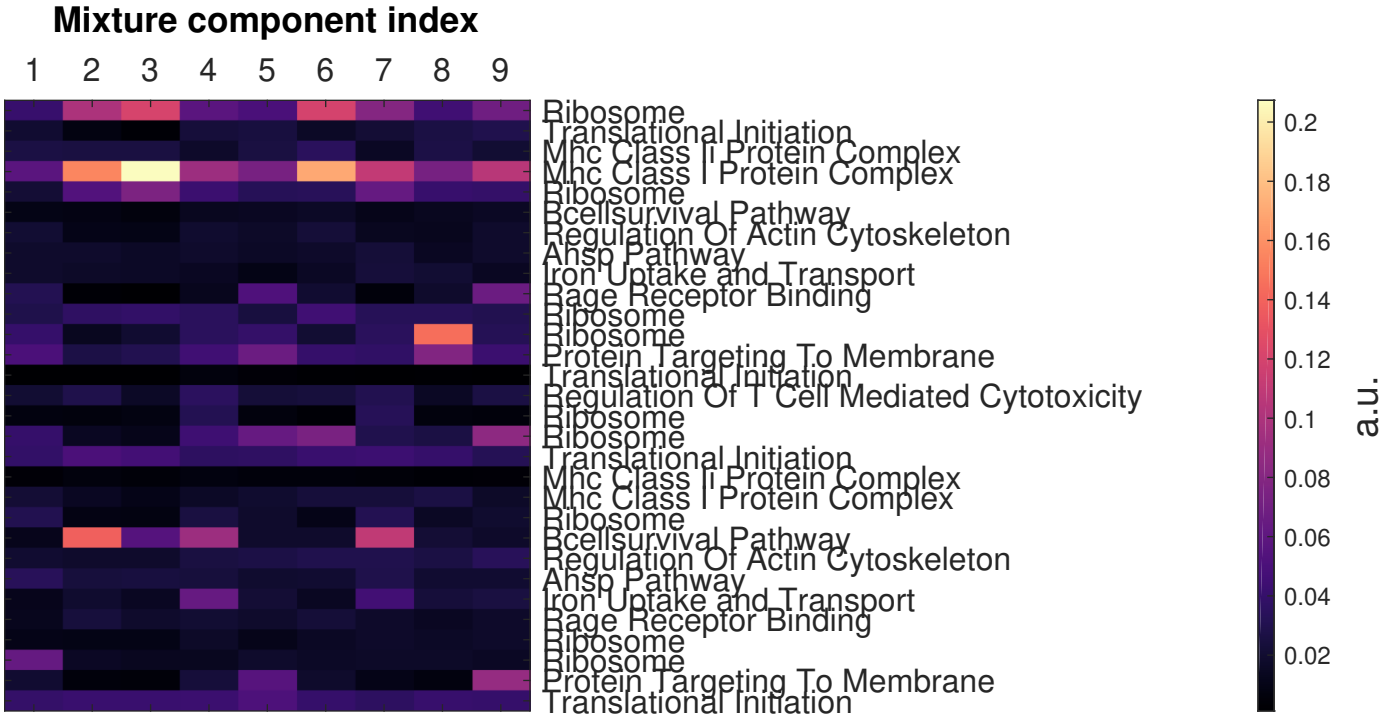




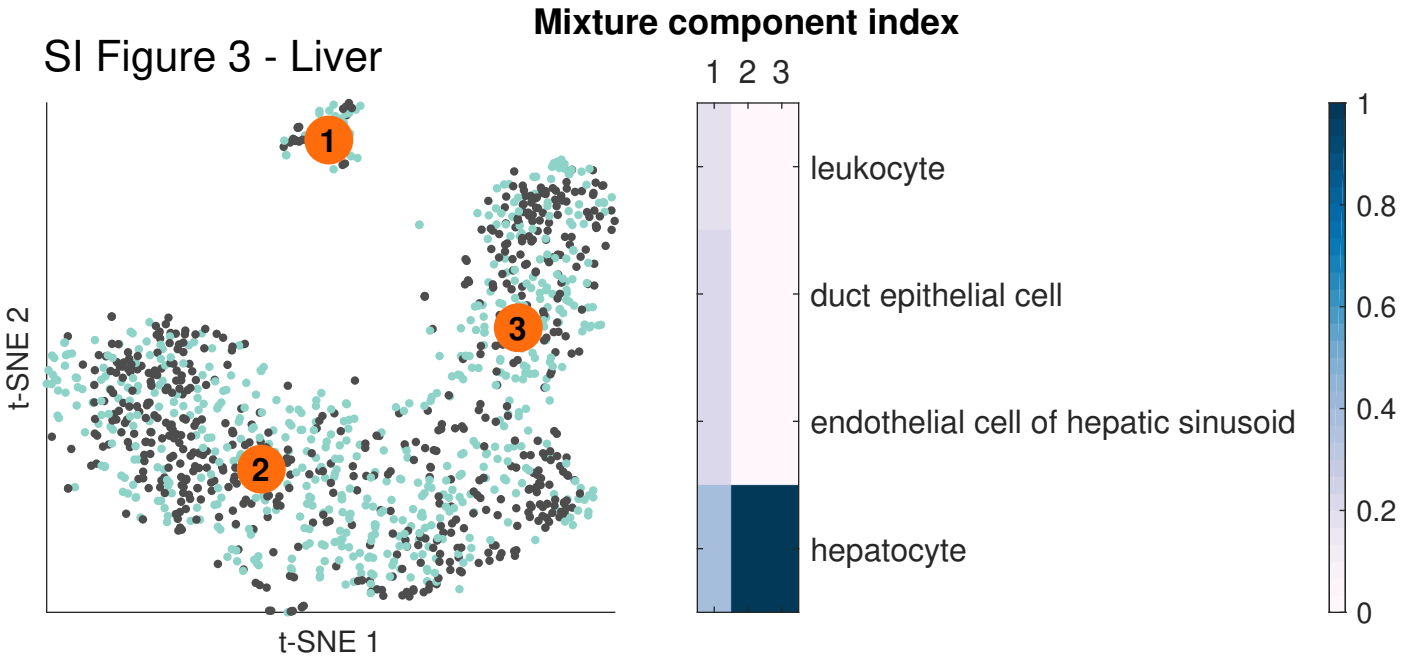
SI Figure 3 - Limb_Muscle

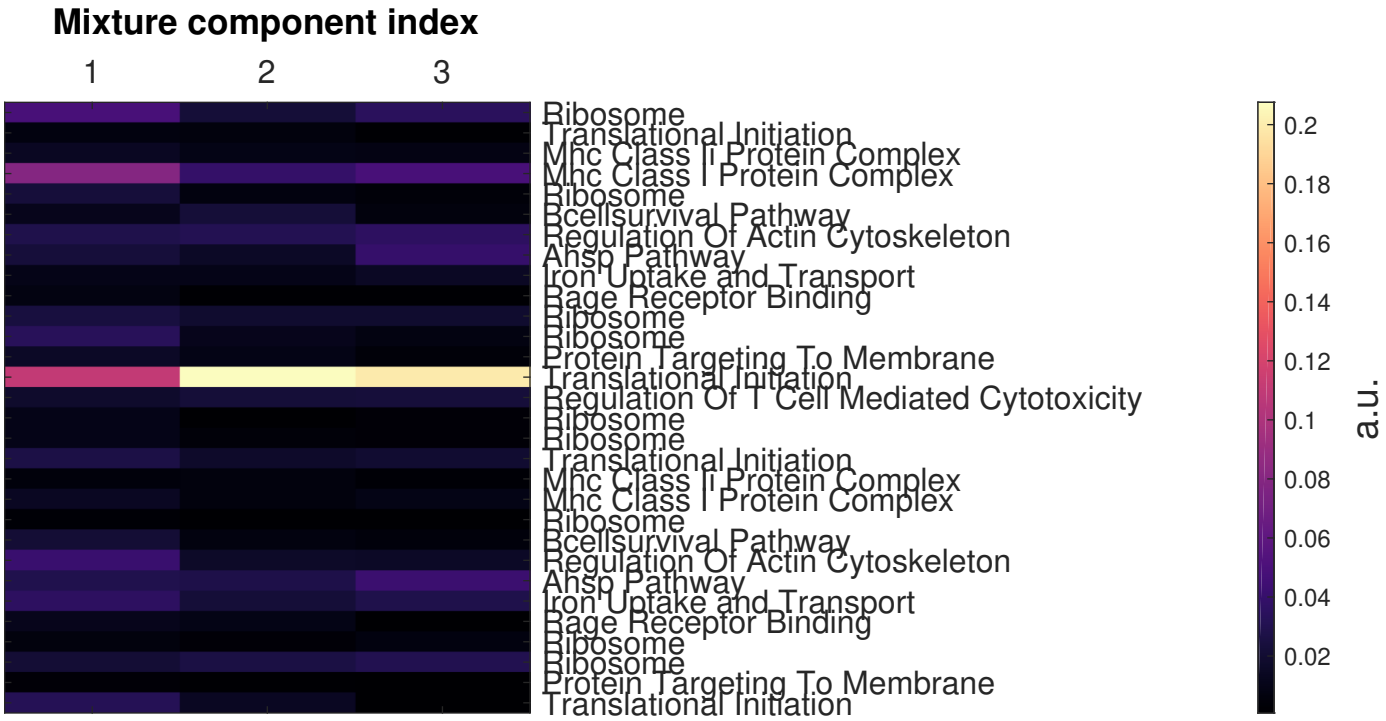




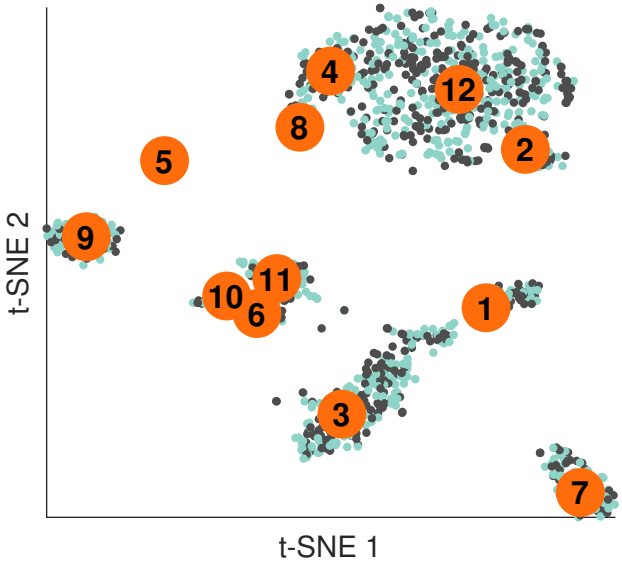


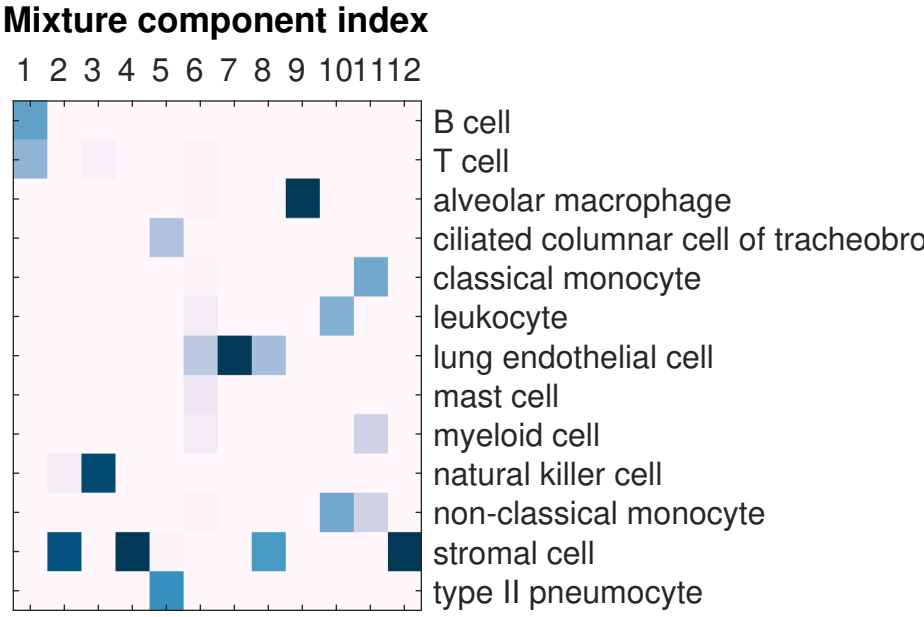
SI Figure 3 - Liver

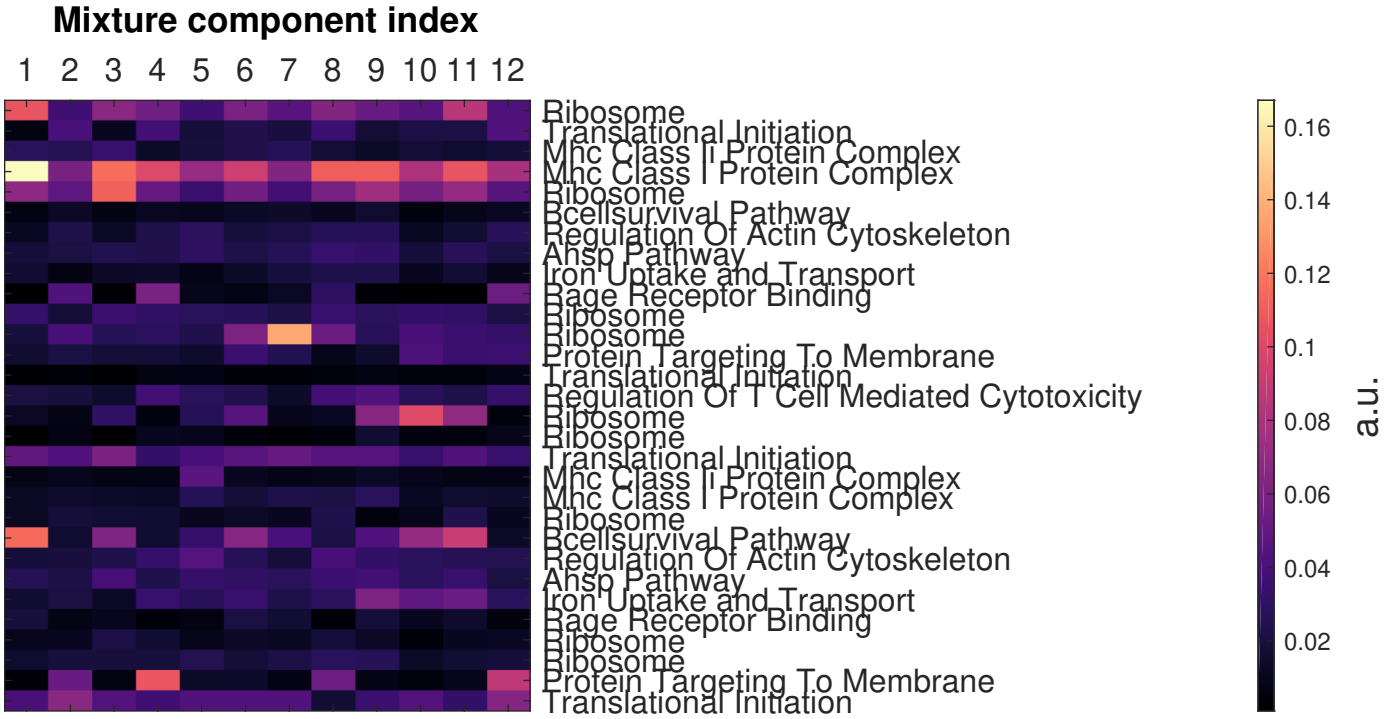




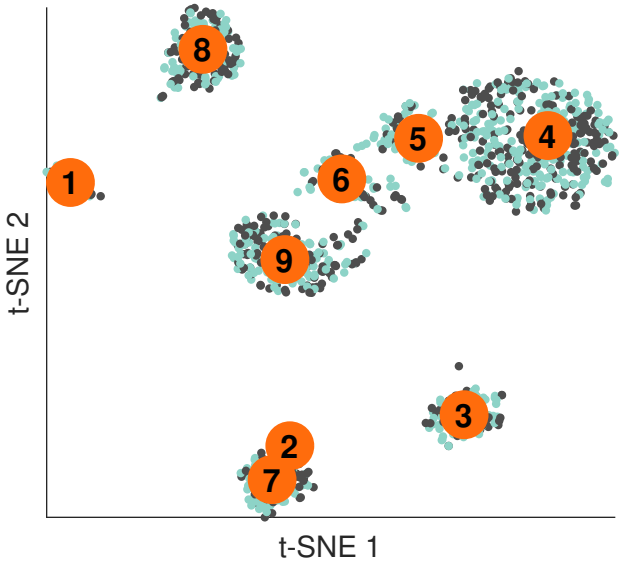
SI Figure 3 - Lung

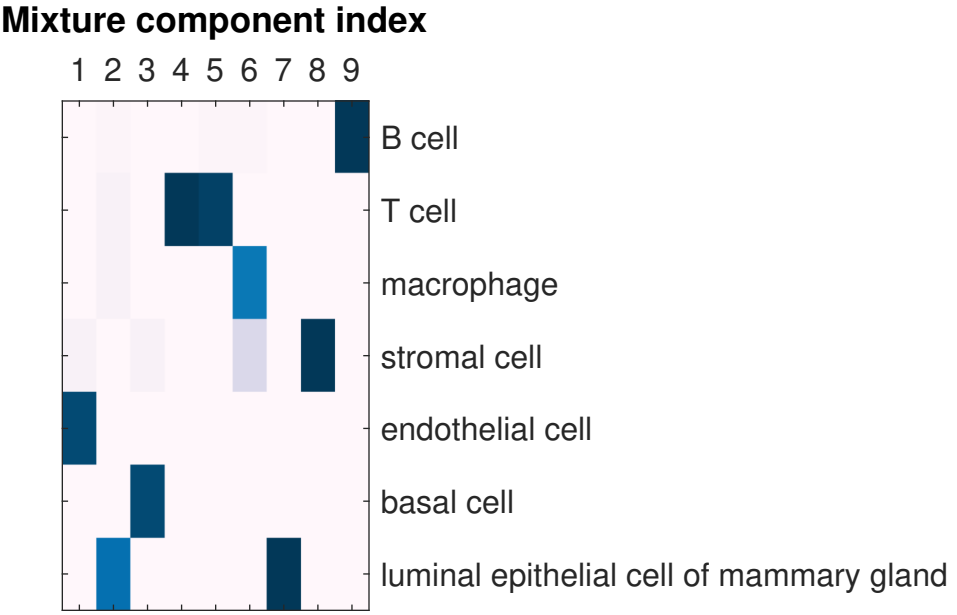


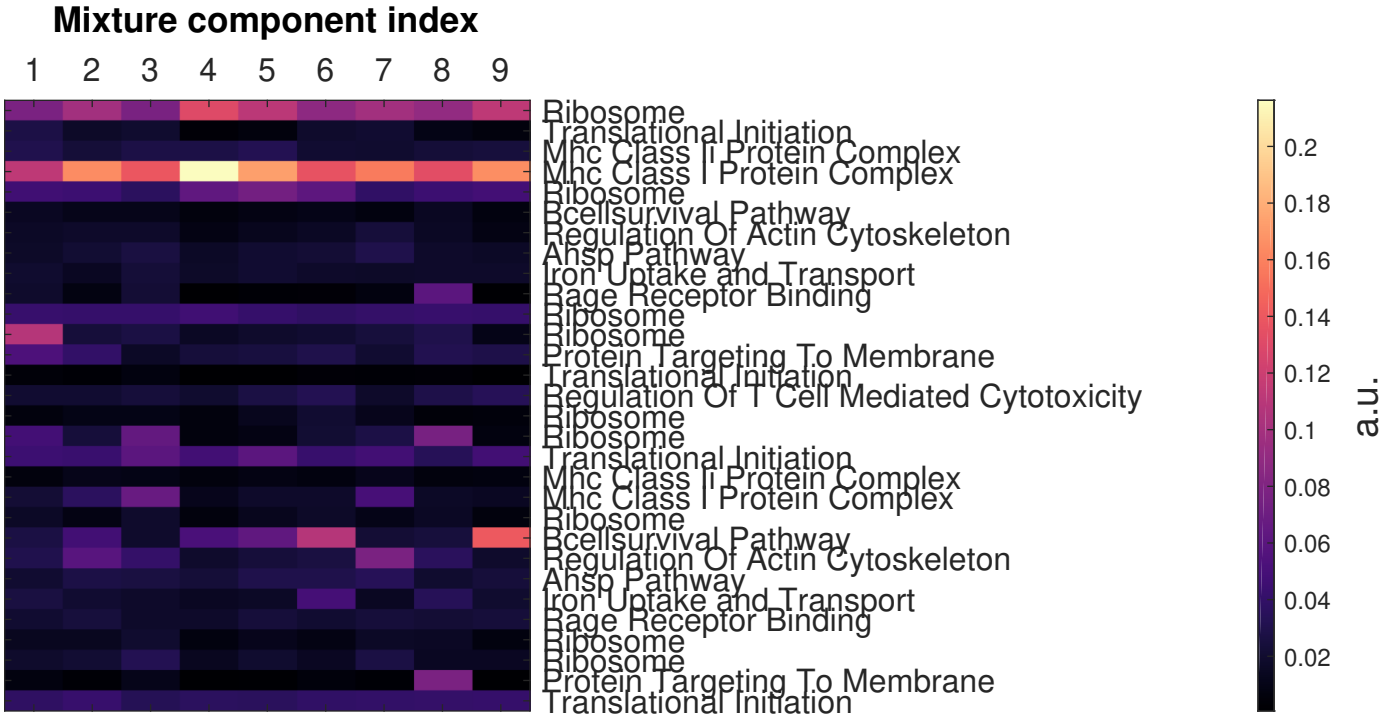




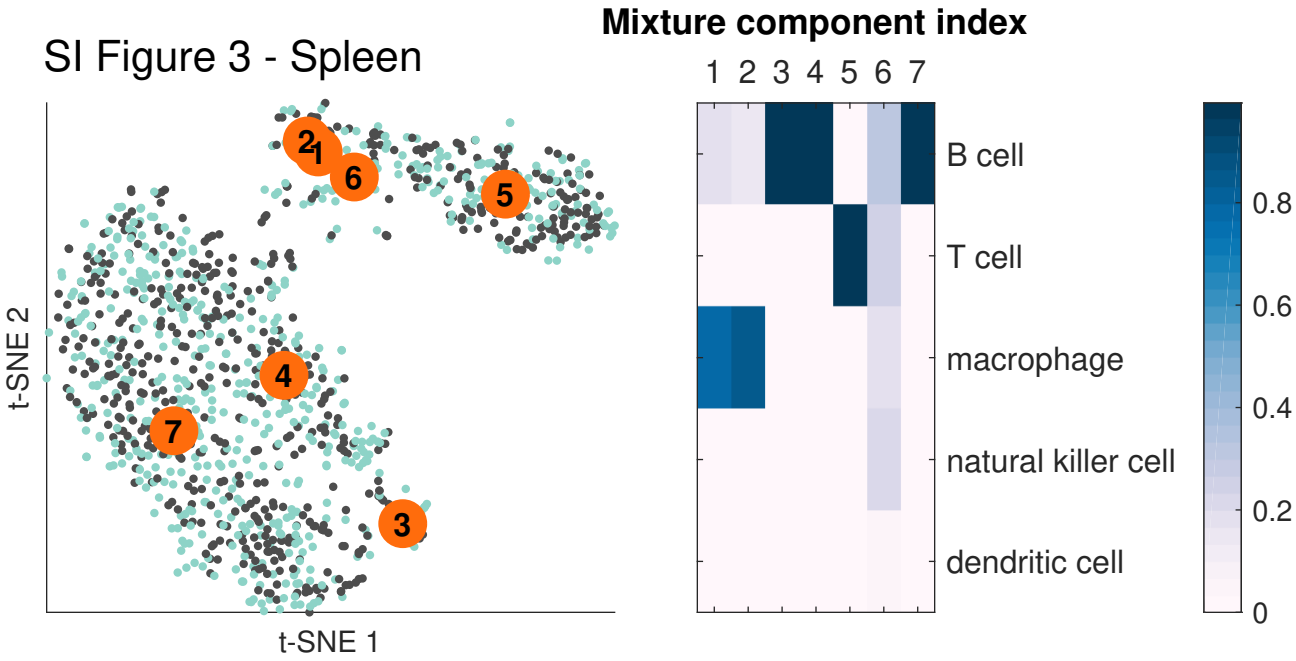
SI Figure 3 - Mammary_Gland

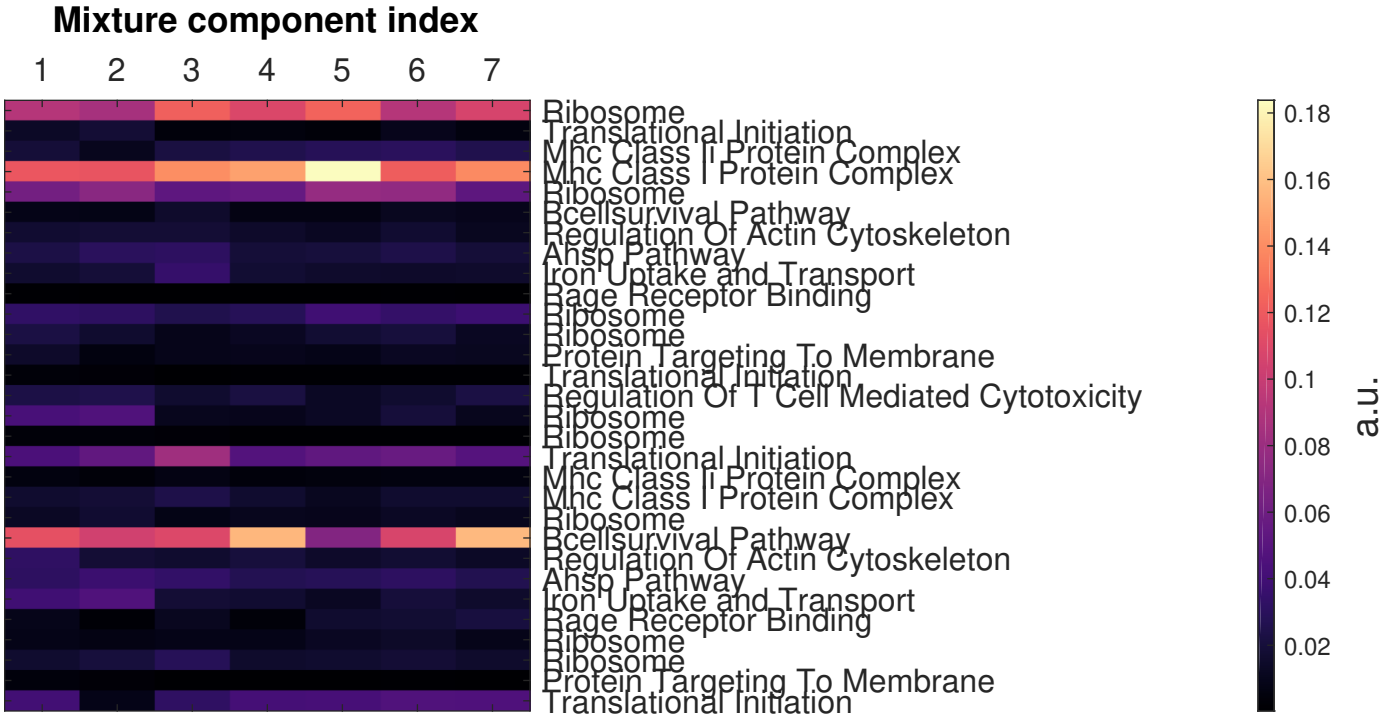




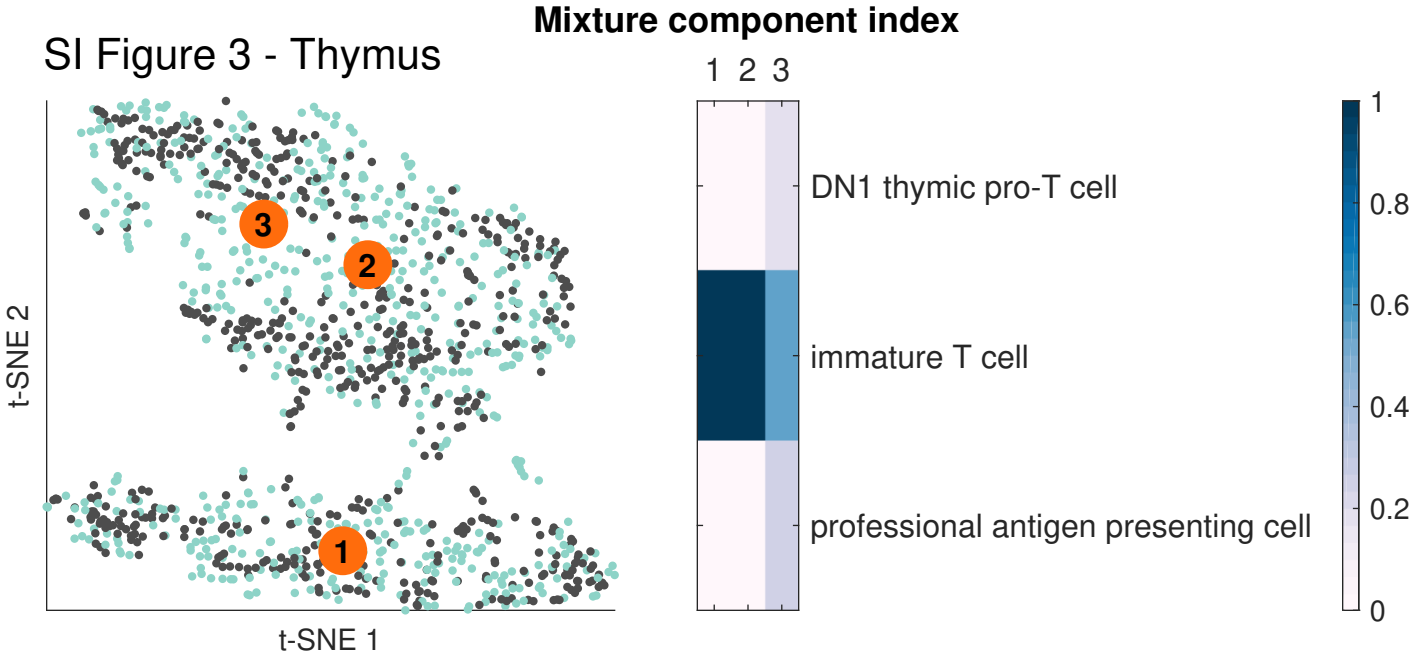


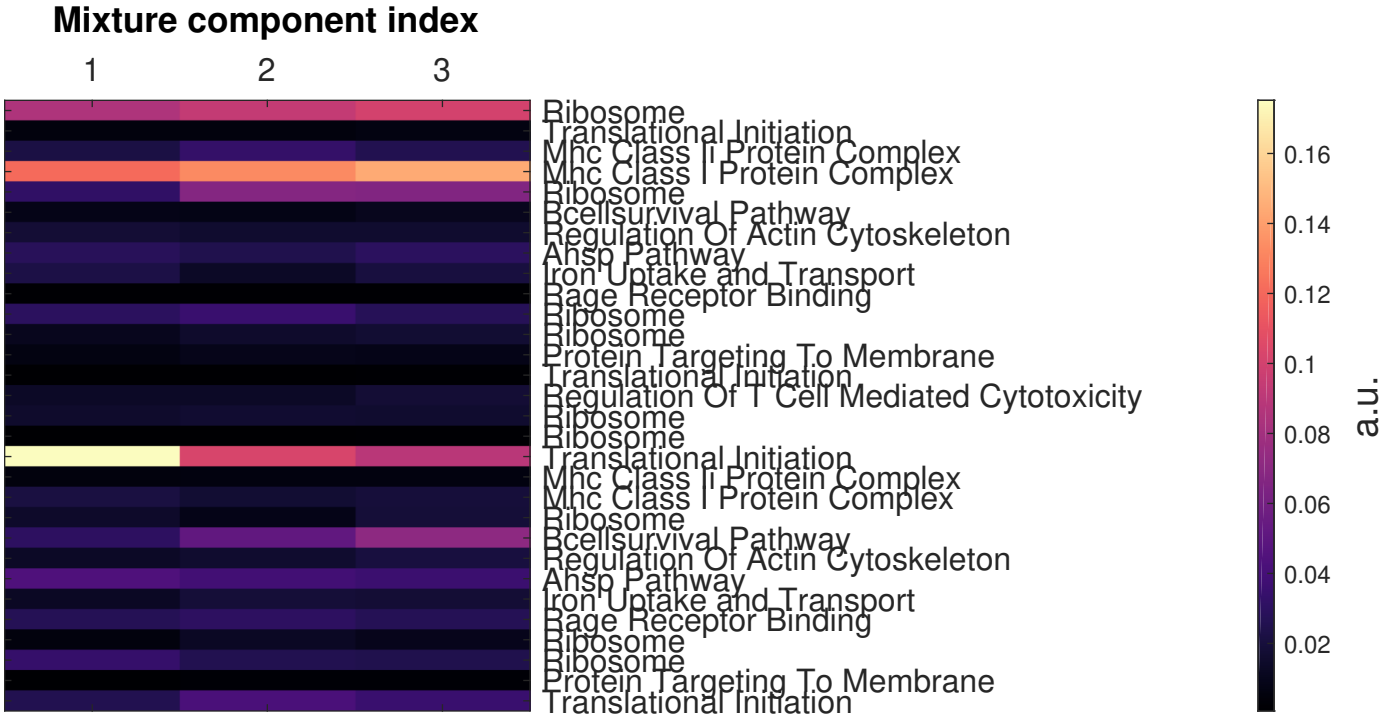
SI Figure 3 - Spleen



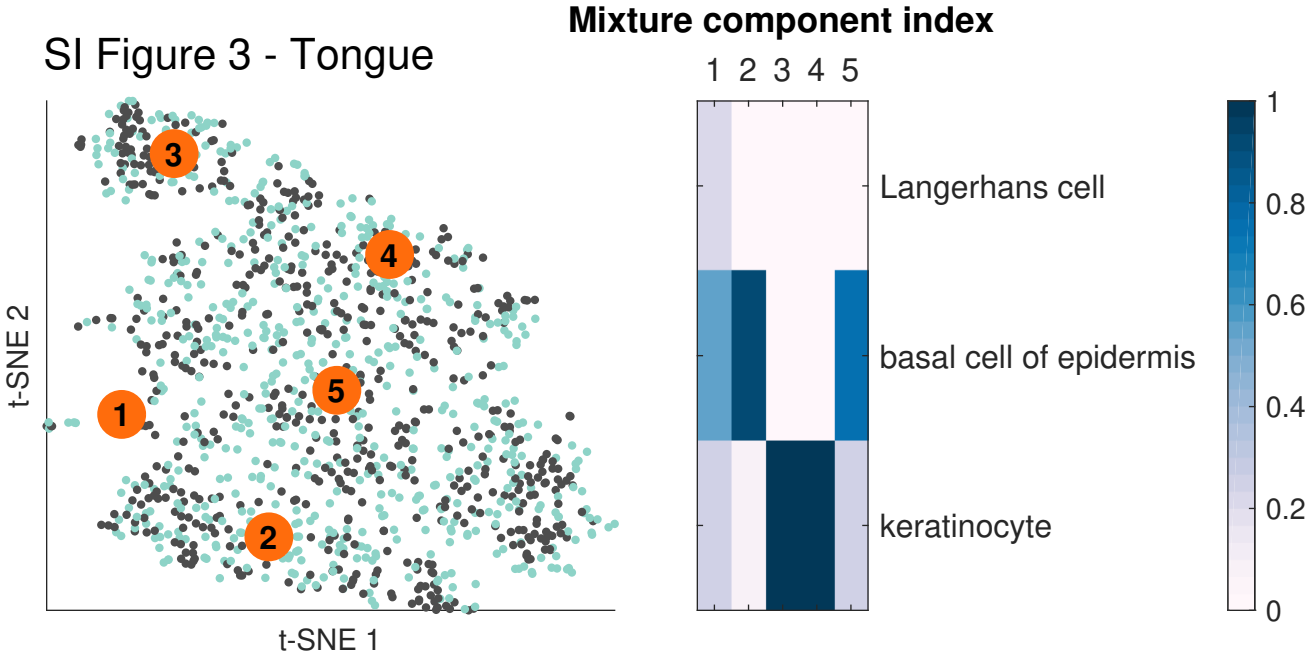


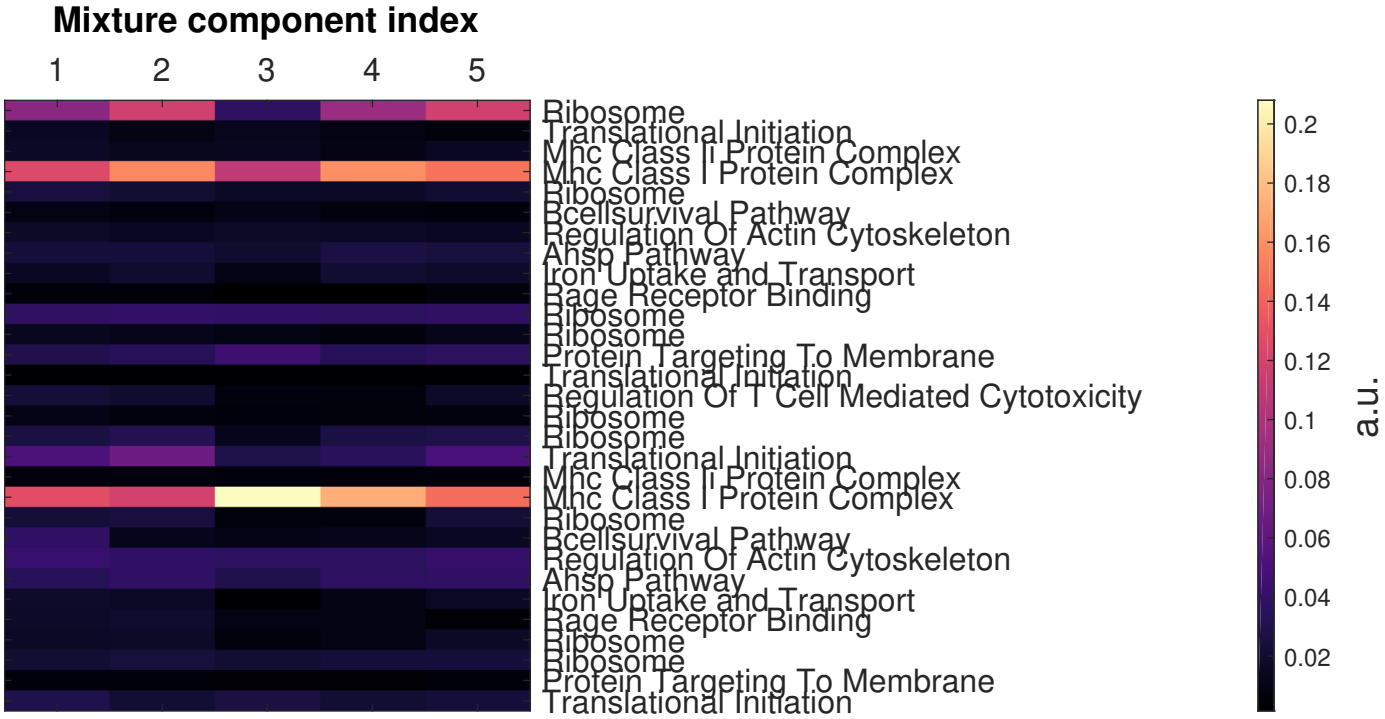
SI Figure 3 - Thymus



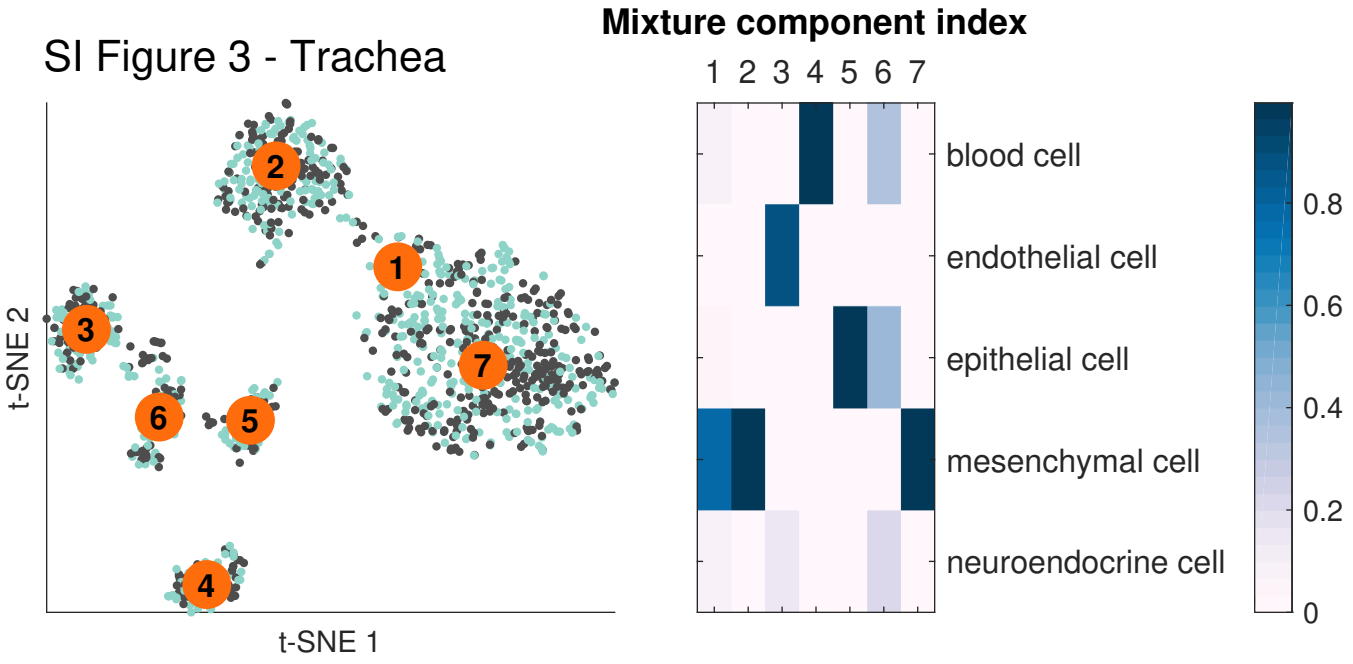


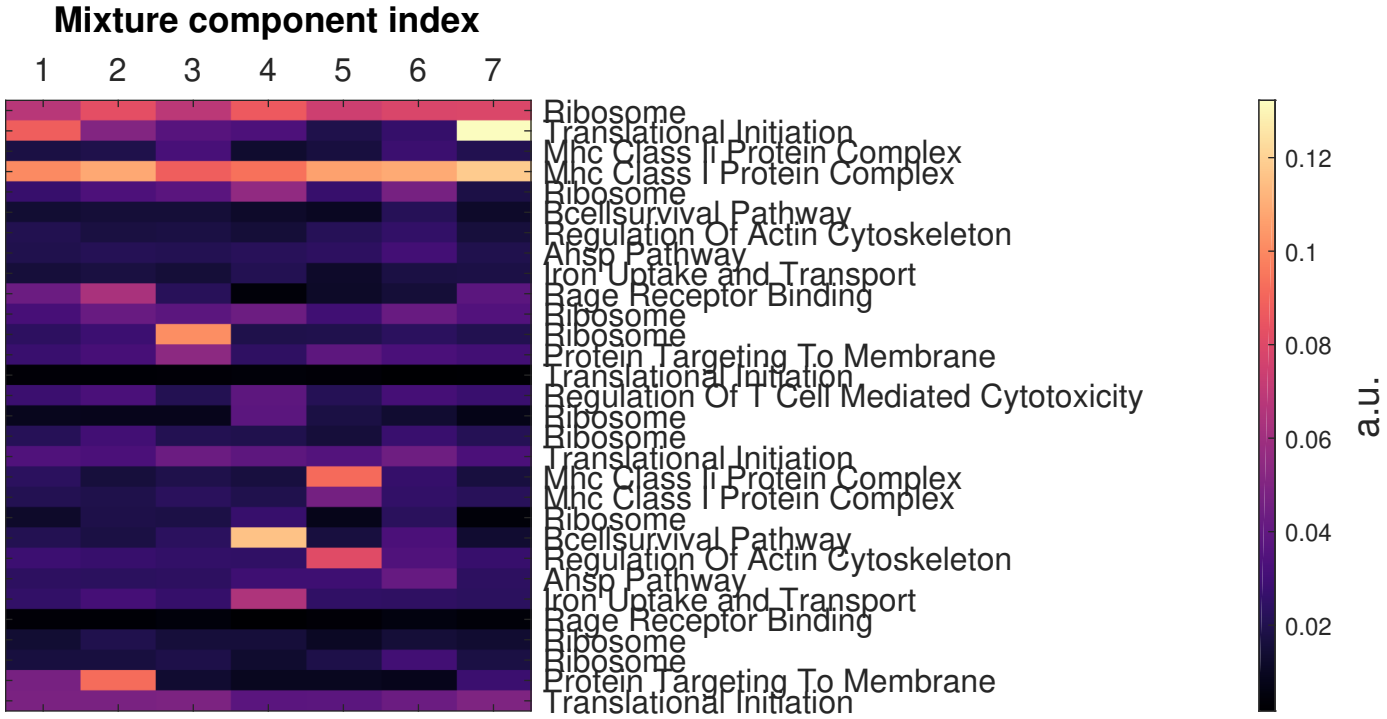
SI Figure 3 - Tongue



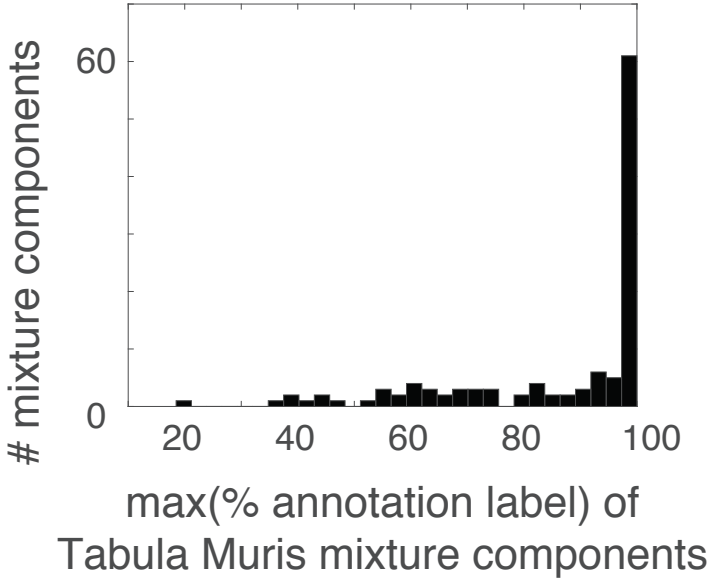


SI Figure 3 - Trachea

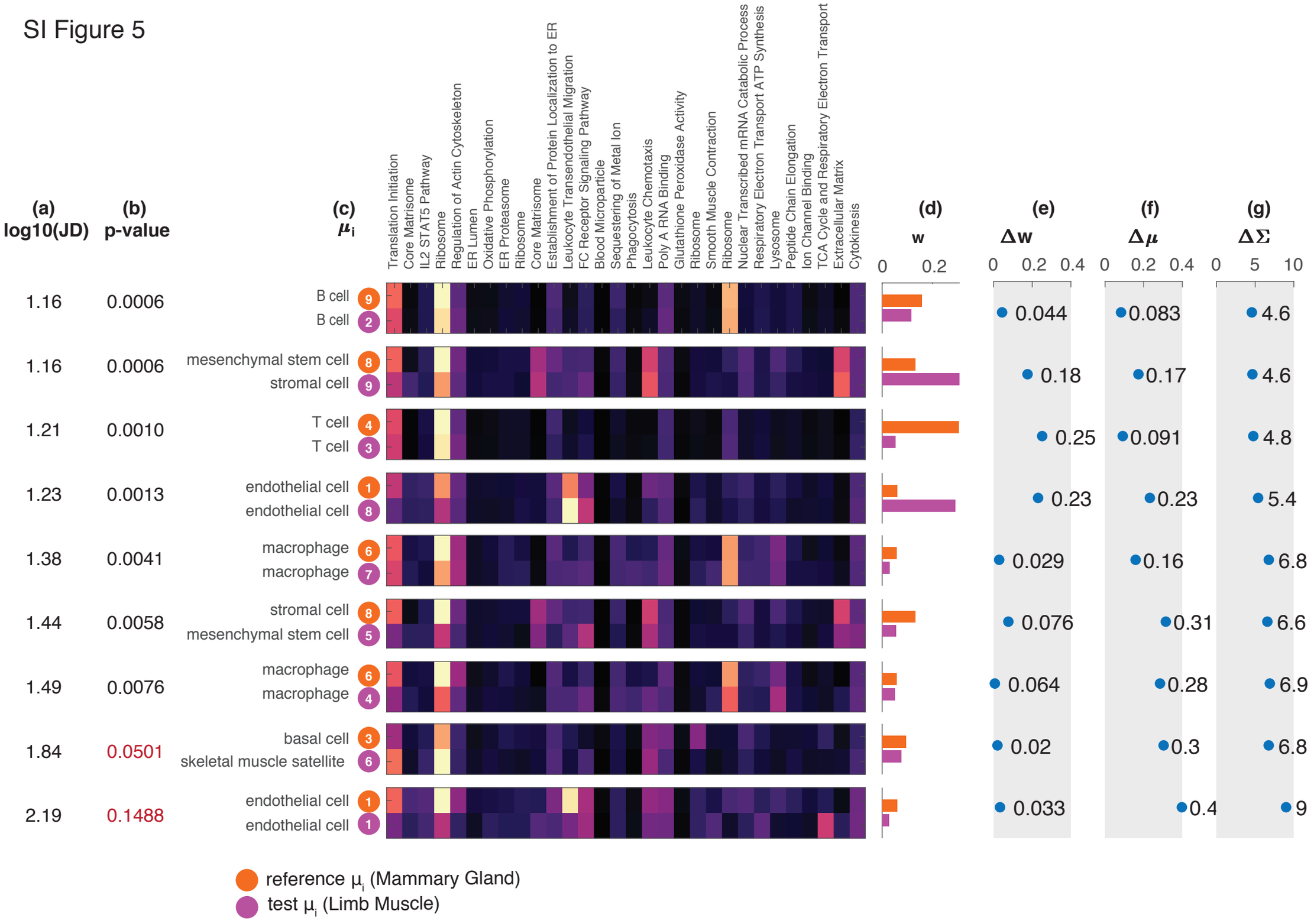




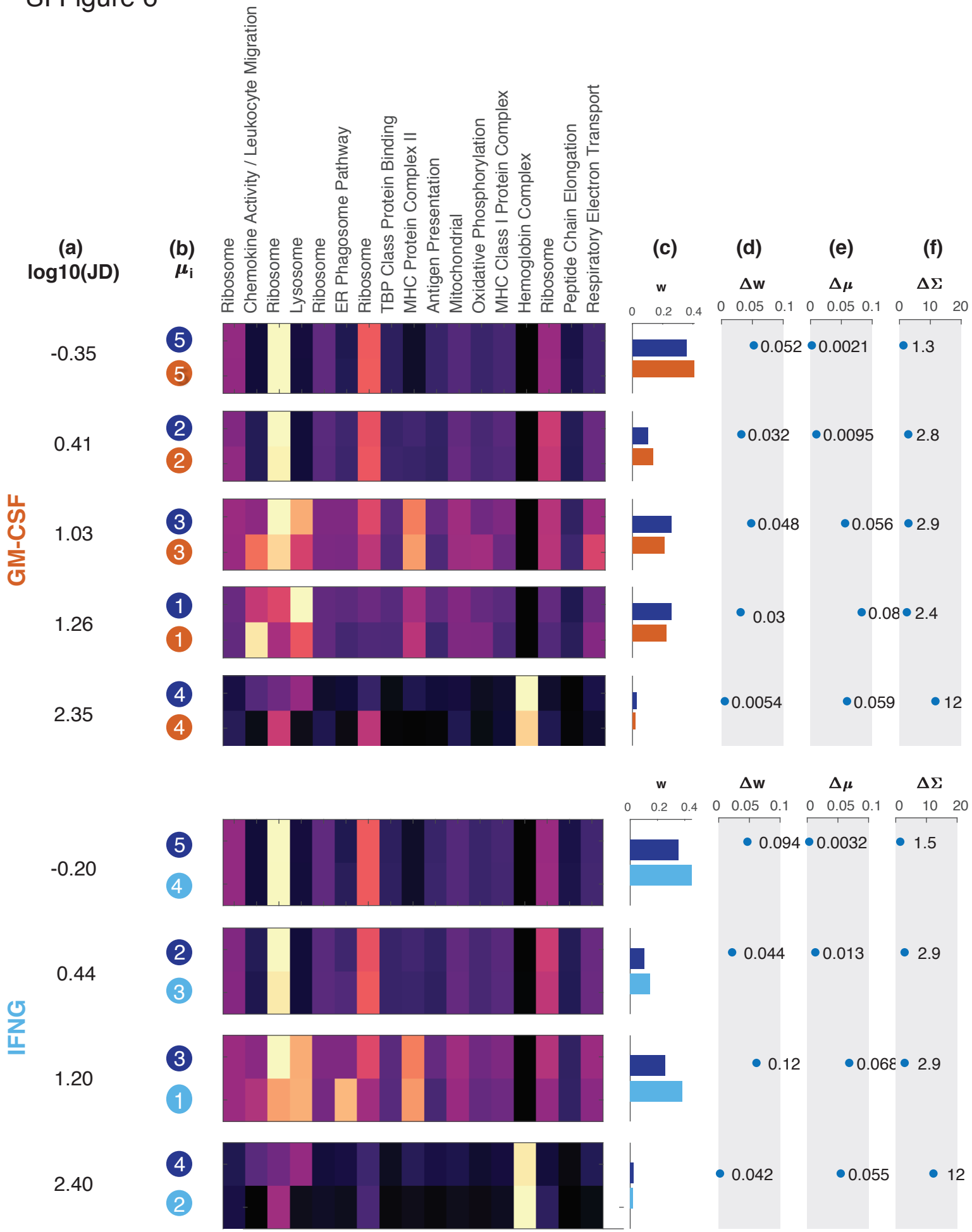
SI Figure 4


mixture components
60
0
20
40
60
80
100
max(% annotation label) of
Tabula Muris mixture components

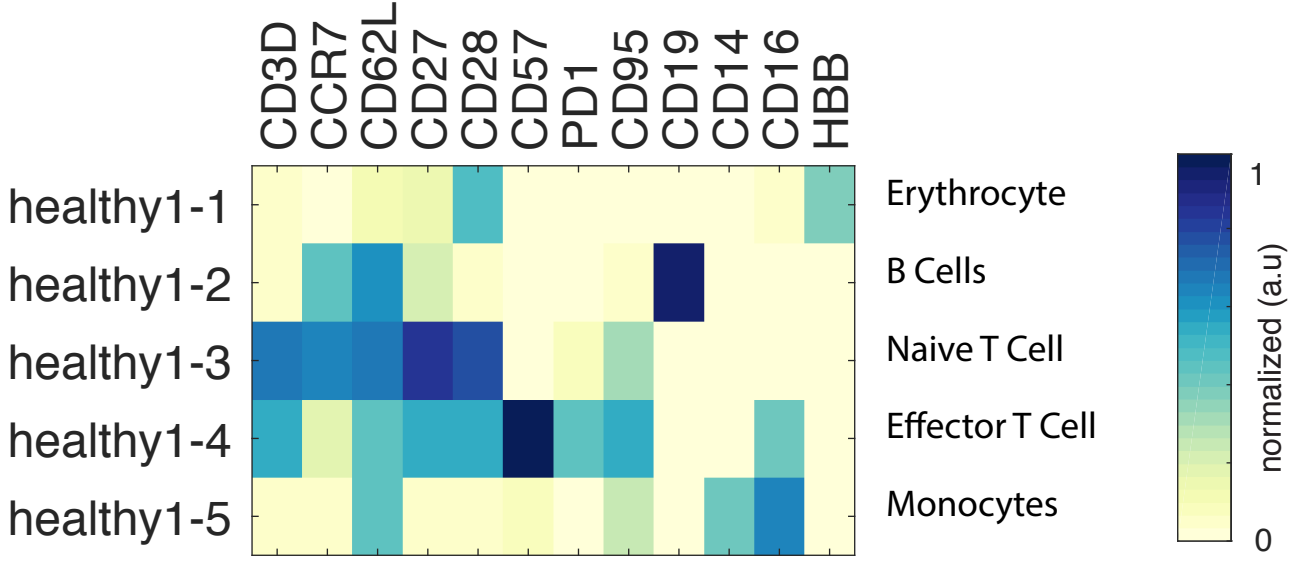
SI Figure 5


(a) log10(JD)
(b) p-value
(c) μ_i
(d) w
(e) Δw
(f) Δμ
(g) ΔΣ
Translation Initiation
Core Matrisome
IL2 STAT5 Pathway
Ribosome
Regulation of Actin Cytoskeleton
ER Lumen
Oxidative Phosphorylation
ER Proteasome
Ribosome
Core Matrisome
Establishment of Protein Localization to ER
Leukocyte Transendothelial Migration
FC Receptor Signaling Pathway
Blood Microparticle
Sequestering of Metal Ion
Phagocytosis
Leukocyte Chemotaxis
Poly A RNA Binding
Glutathione Peroxidase Activity
Ribosome
Smooth Muscle Contraction
Ribosome
Nuclear Transcribed mRNA Catabolic Process
Respiratory Electron Transport ATP Synthesis
Lysosome
Peptide Chain Elongation
Ion Channel Binding
TCA Cycle and Respiratory Electron Transport
Extracellular Matrix
Cytokinesis
0
0.2
0
0.2
0.4
0
0.2
0.4
0
5
10
1.16
0.0006
B cell 9
B cell 2
0.044
0.083
4.6
1.16
0.0006
mesenchymal stem cell 8
stromal cell 9
0.18
0.17
4.6
1.21
0.0010
T cell 4
T cell 3
0.25
0.091
4.8
1.23
0.0013
endothelial cell 1
endothelial cell 8
0.23
0.23
5.4
1.38
0.0041
macrophage 6
macrophage 7
0.029
0.16
6.8
1.44
0.0058
stromal cell 8
mesenchymal stem cell 5
0.076
0.31
6.6
1.49
0.0076
macrophage 6
macrophage 4
0.064
0.28
6.9
1.84
0.0501
basal cell 3
skeletal muscle satellite 6
0.02
0.3
6.8
2.19
0.1488
endothelial cell 1
endothelial cell 1
0.033
0.4
9
reference μi (Mammary Gland)
test μi (Limb Muscle)

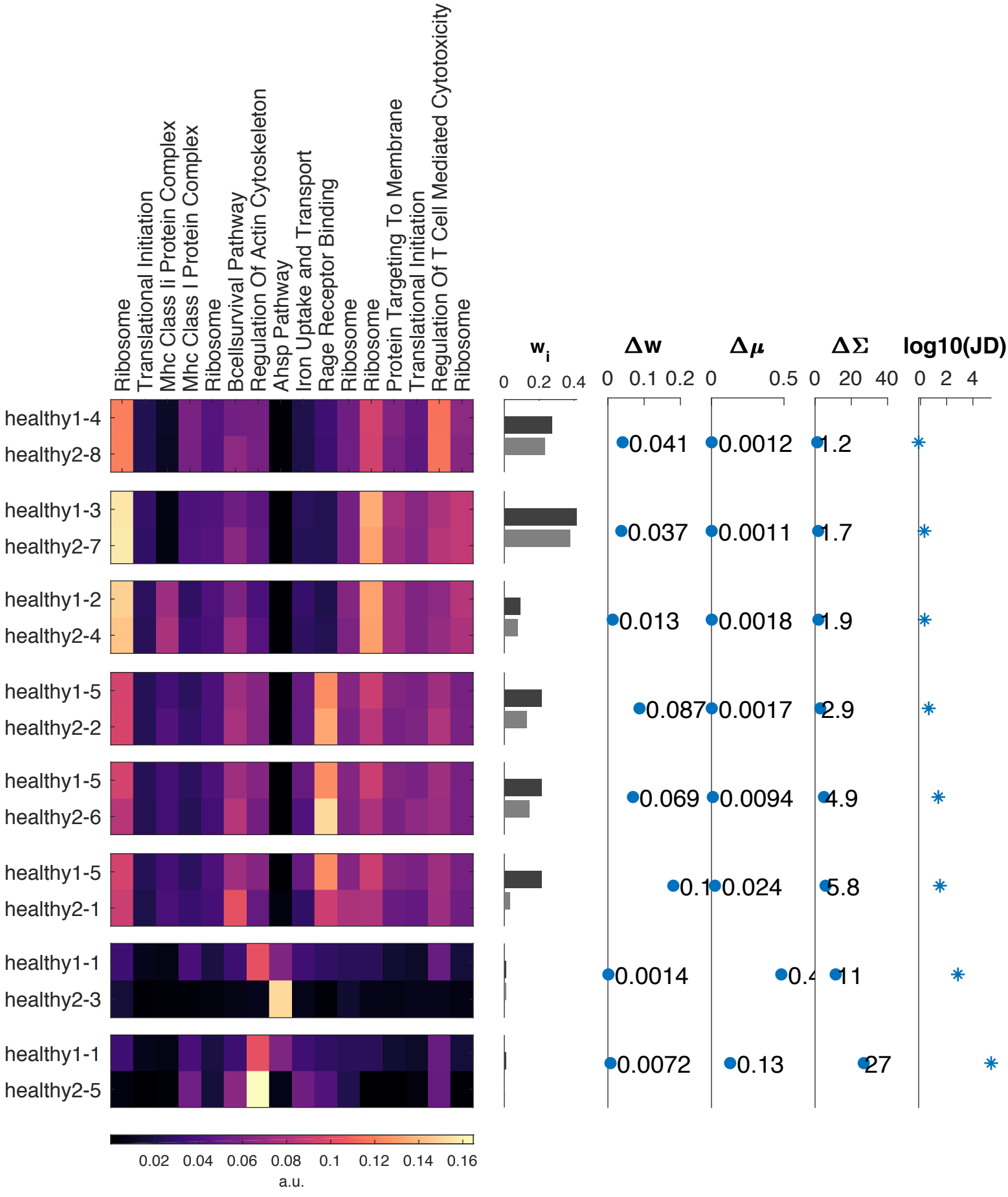
SI Figure 6


Ribosome
Chemokine Activity / Leukocyte Migration
Ribosome
Lysosome
Ribosome
ER Phagosome Pathway
Ribosome
TBP Class Protein Binding
MHC Protein Complex II
Antigen Presentation
Mitochondrial
Oxidative Phosphorylation
MHC Class I Protein Complex
Hemoglobin Complex
Ribosome
Peptide Chain Elongation
Respiratory Electron Transport
(a)
log10(JD)
(b)
μ_i
(c)
w
(d)
Δw
(e)
Δμ
(f)
ΔΣ
GM-CSF
-0.35
0.41
1.03
1.26
2.35
0.052
0.0021
1.3
0.032
0.0095
2.8
0.048
0.056
2.9
0.03
0.08
2.4
0.0054
0.059
12
IFNG
-0.20
0.44
1.20
2.40
0.094
0.0032
1.5
0.044
0.013
2.9
0.12
0.068
2.9
0.042
0.055
12

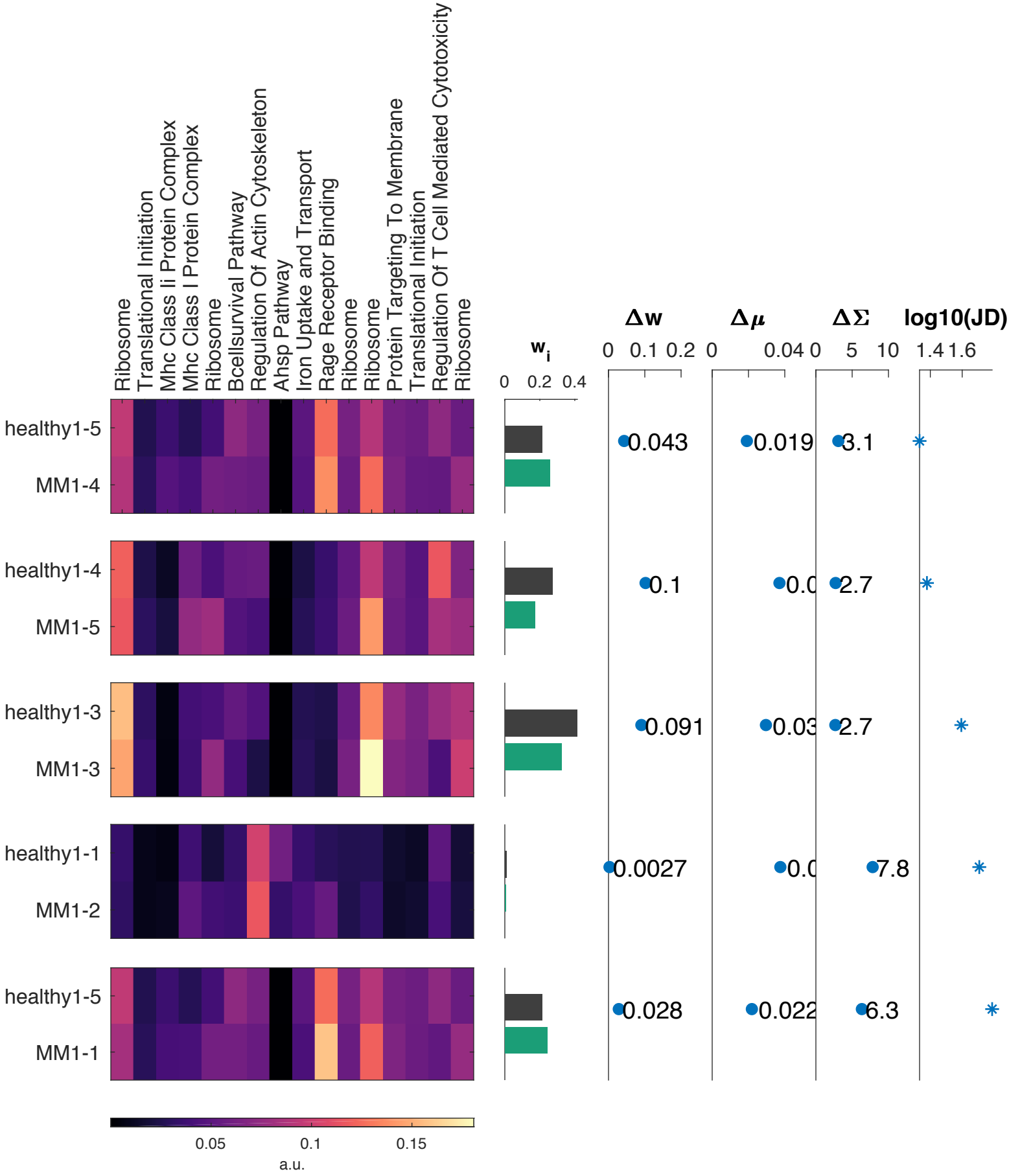
SI Figure 7


CD3D
CCR7
CD62L
CD27
CD28
CD57
PD1
CD95
CD19
CD14
CD16
HBB
healthy1-1
healthy1-2
healthy1-3
healthy1-4
healthy1-5
Erythrocyte
B Cells
Naive T Cell
Effector T Cell
Monocytes
1
0
normalized (a.u)

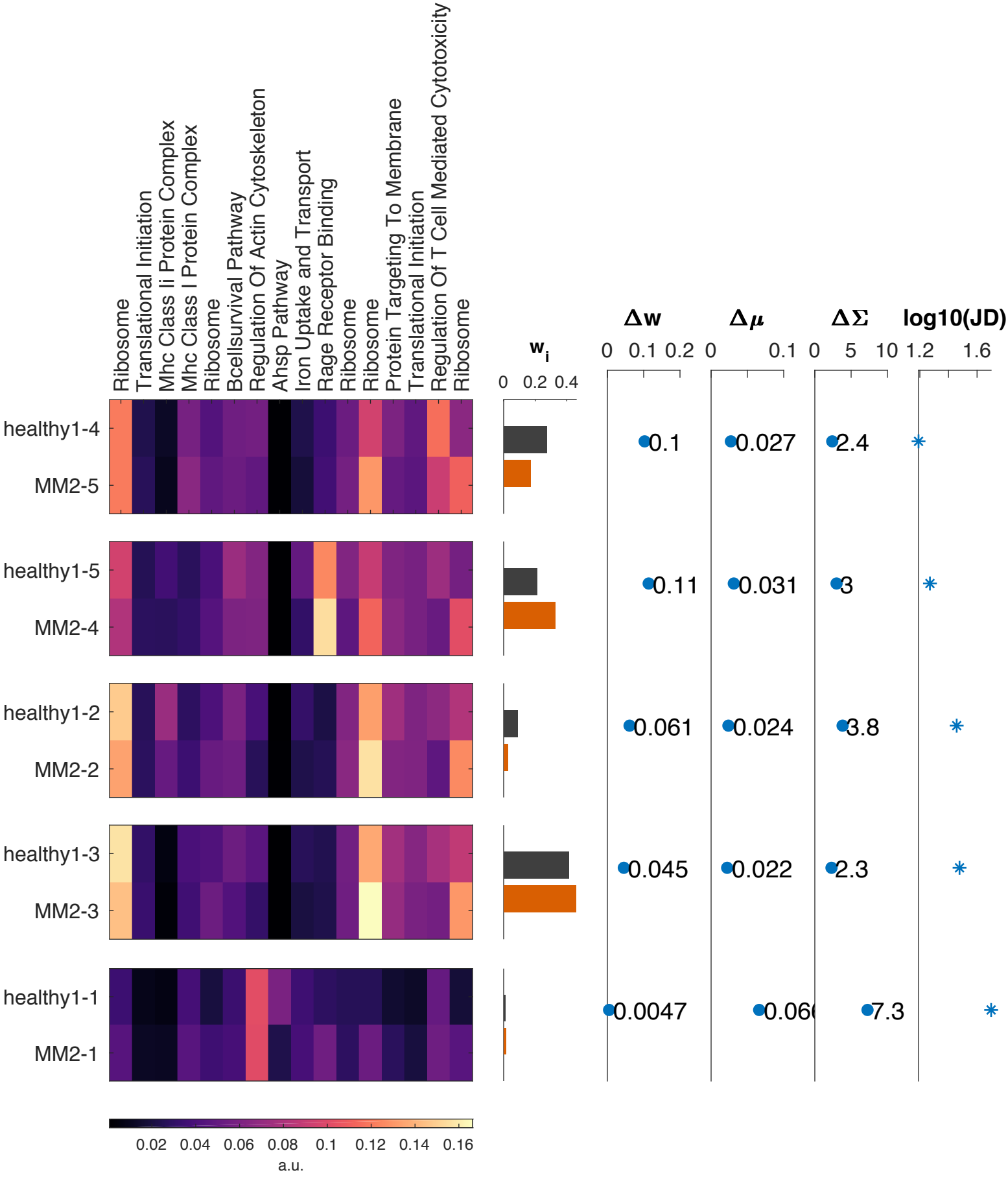
SI Figure 8


Ribosome
Translational Initiation
Mhc Class Ii Protein Complex
Mhc Class I Protein Complex
Ribosome
Bcellsurvival Pathway
Regulation Of Actin Cytoskeleton
Ahsp Pathway
Iron Uptake and Transport
Rage Receptor Binding
Ribosome
Ribosome
Protein Targeting To Membrane
Translational Initiation
Regulation Of T Cell Mediated Cytotoxicity
Ribosome
healthy1-4
healthy2-8
healthy1-3
healthy2-7
healthy1-2
healthy2-4
healthy1-5
healthy2-2
healthy1-5
healthy2-6
healthy1-5
healthy2-1
healthy1-1
healthy2-3
healthy1-1
healthy2-5
0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16
a.u.
w_i
0 0.2 0.4
Δw
0 0.1 0.2
Δμ
0 0.5
ΔΣ
0 20 40
log10(JD)
0 2 4
0.041
0.0012
1.2
0.037
0.0011
1.7
0.013
0.0018
1.9
0.087
0.0017
2.9
0.069
0.0094
4.9
0.1
0.024
5.8
0.0014
11
0.0072
0.13
27

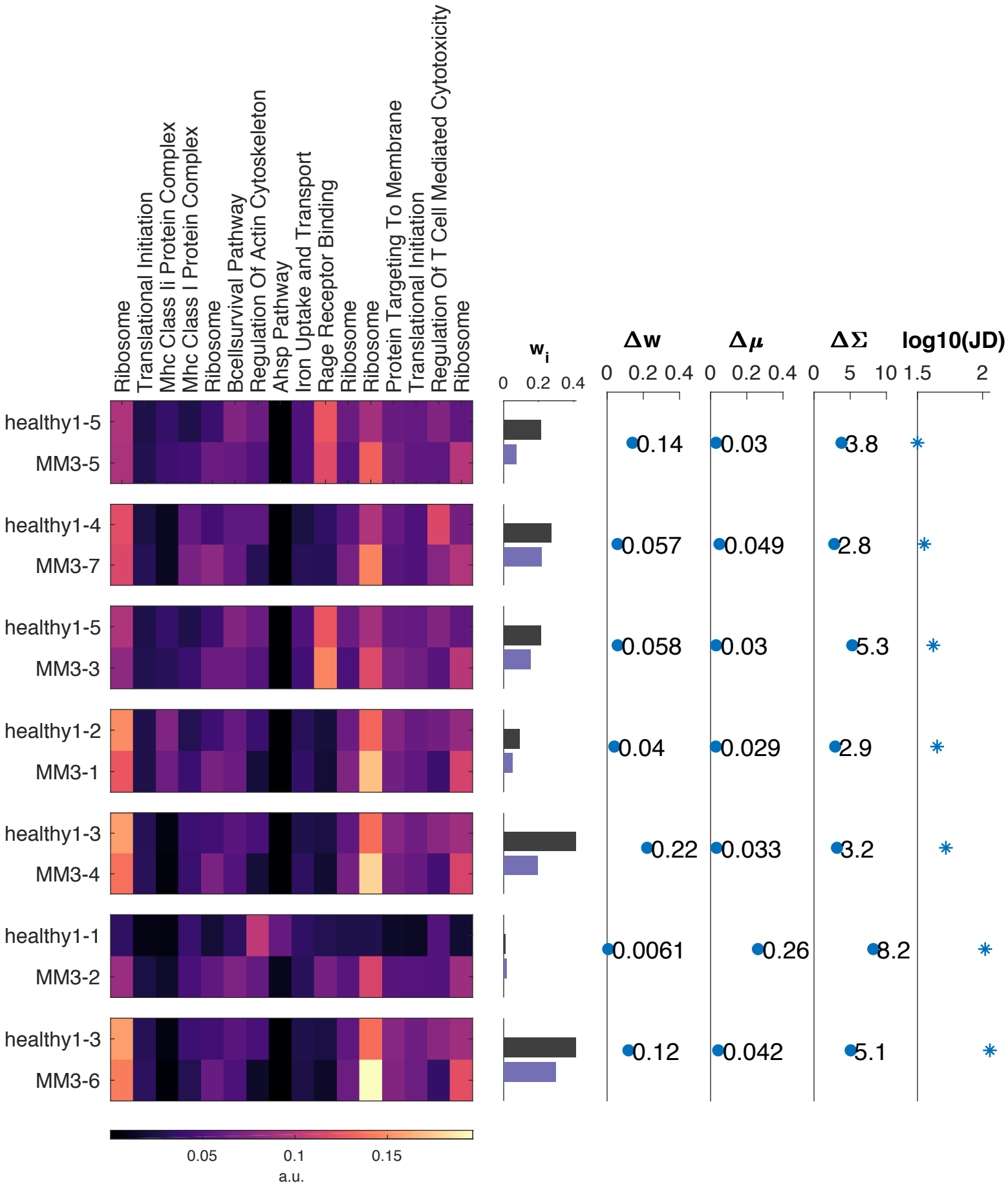
SI Figure 9


Ribosome
Translational Initiation
Mhc Class Ii Protein Complex
Mhc Class I Protein Complex
Ribosome
Bcellsurvival Pathway
Regulation Of Actin Cytoskeleton
Ahsp Pathway
Iron Uptake and Transport
Rage Receptor Binding
Ribosome
Ribosome
Protein Targeting To Membrane
Translational Initiation
Regulation Of T Cell Mediated Cytotoxicity
Ribosome
healthy1-5
MM1-4
healthy1-4
MM1-5
healthy1-3
MM1-3
healthy1-1
MM1-2
healthy1-5
MM1-1
0.05
0.1
0.15
a.u.
w_i
0
0.2
0.4
Δw
Δμ
ΔΣ
log10(JD)
0
0.1
0.2
0
0.04
0
5
10
1.4
1.6
0.043
0.019
3.1
0.1
0.0
2.7
0.091
0.03
2.7
0.0027
0.0
7.8
0.028
0.022
6.3

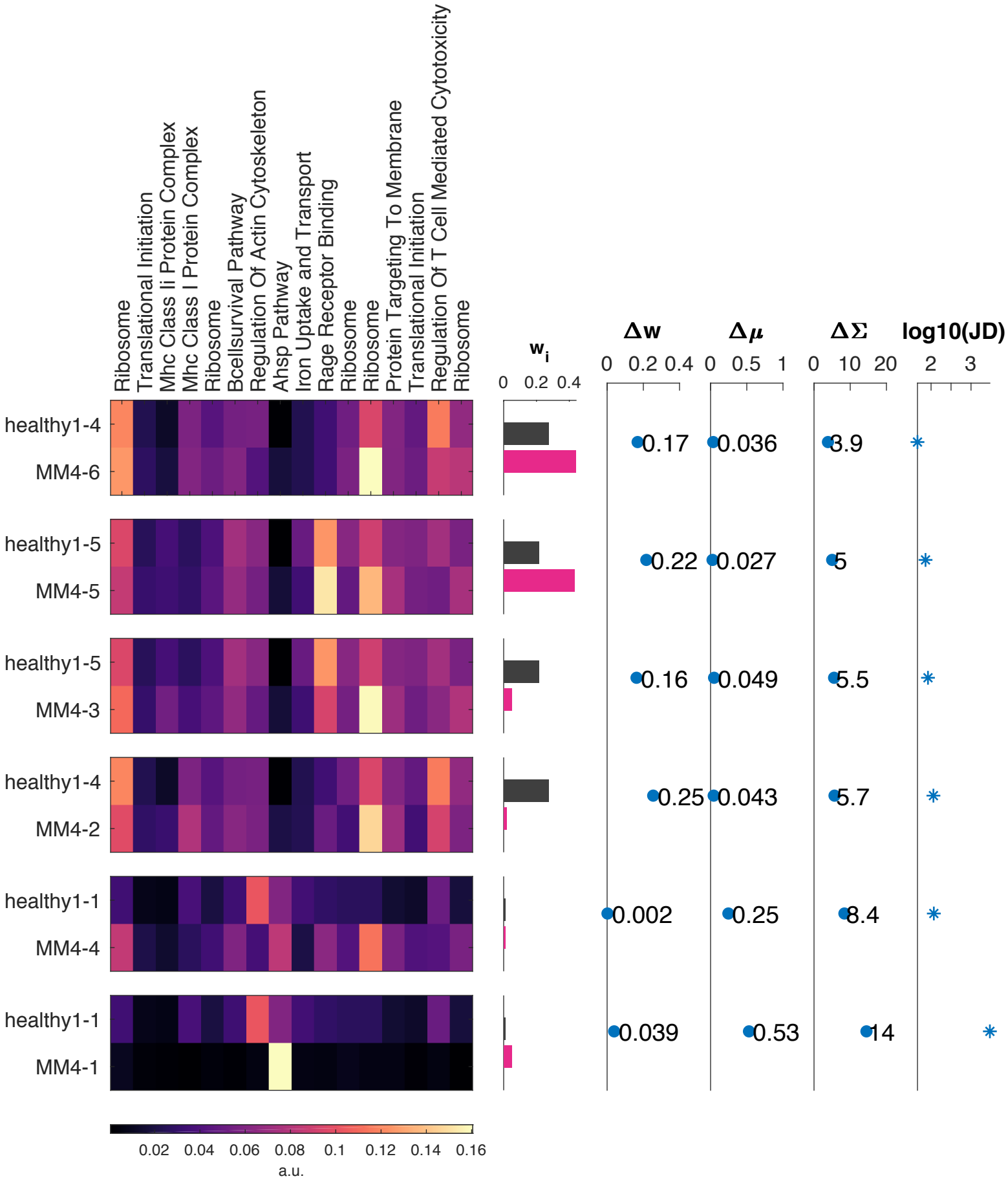
SI Figure 10


Ribosome
Translational Initiation
Mhc Class Ii Protein Complex
Mhc Class I Protein Complex
Ribosome
Bcellsurvival Pathway
Regulation Of Actin Cytoskeleton
Ahsp Pathway
Iron Uptake and Transport
Rage Receptor Binding
Ribosome
Ribosome
Protein Targeting To Membrane
Translational Initiation
Regulation Of T Cell Mediated Cytotoxicity
Ribosome
healthy1-4
MM2-5
healthy1-5
MM2-4
healthy1-2
MM2-2
healthy1-3
MM2-3
healthy1-1
MM2-1
0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16
a.u.
w_i
0 0.2 0.4
Δw
0 0.1 0.2
Δμ
0 0.1
ΔΣ
0 5 10
log10(JD)
1.2 1.6
0.1
0.027
2.4
0.11
0.031
3
0.061
0.024
3.8
0.045
0.022
2.3
0.0047
0.06
7.3

SI Figure 11


Ribosome
Translational Initiation
Mhc Class Ii Protein Complex
Mhc Class I Protein Complex
Ribosome
Bcellsurvival Pathway
Regulation Of Actin Cytoskeleton
Ahsp Pathway
Iron Uptake and Transport
Rage Receptor Binding
Ribosome
Ribosome
Protein Targeting To Membrane
Translational Initiation
Regulation Of T Cell Mediated Cytotoxicity
Ribosome
healthy1-5
MM3-5
healthy1-4
MM3-7
healthy1-5
MM3-3
healthy1-2
MM3-1
healthy1-3
MM3-4
healthy1-1
MM3-2
healthy1-3
MM3-6
0.05
0.1
0.15
a.u.
w_i
0 0.2 0.4
Δw
0 0.2 0.4
Δμ
0 0.2 0.4
ΔΣ
0 5 10
log10(JD)
1.5 2
0.14
0.03
3.8
0.057
0.049
2.8
0.058
0.03
5.3
0.04
0.029
2.9
0.22
0.033
3.2
0.0061
0.26
8.2
0.12
0.042
5.1

SI Figure 12


Ribosome
Translational Initiation
Mhc Class Ii Protein Complex
Mhc Class I Protein Complex
Ribosome
Bcellsurvival Pathway
Regulation Of Actin Cytoskeleton
Ahsp Pathway
Iron Uptake and Transport
Rage Receptor Binding
Ribosome
Ribosome
Protein Targeting To Membrane
Translational Initiation
Regulation Of T Cell Mediated Cytotoxicity
Ribosome
healthy1-4
MM4-6
healthy1-5
MM4-5
healthy1-5
MM4-3
healthy1-4
MM4-2
healthy1-1
MM4-4
healthy1-1
MM4-1
0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16
a.u.
w_i
0 0.2 0.4
Δw
0 0.2 0.4
Δμ
0 0.5 1
ΔΣ
0 10 20
log10(JD)
2 3
0.17
0.036
3.9
0.22
0.027
5
0.16
0.049
5.5
0.25
0.043
5.7
0.002
0.25
8.4
0.039
0.53
14

SI Figure 13

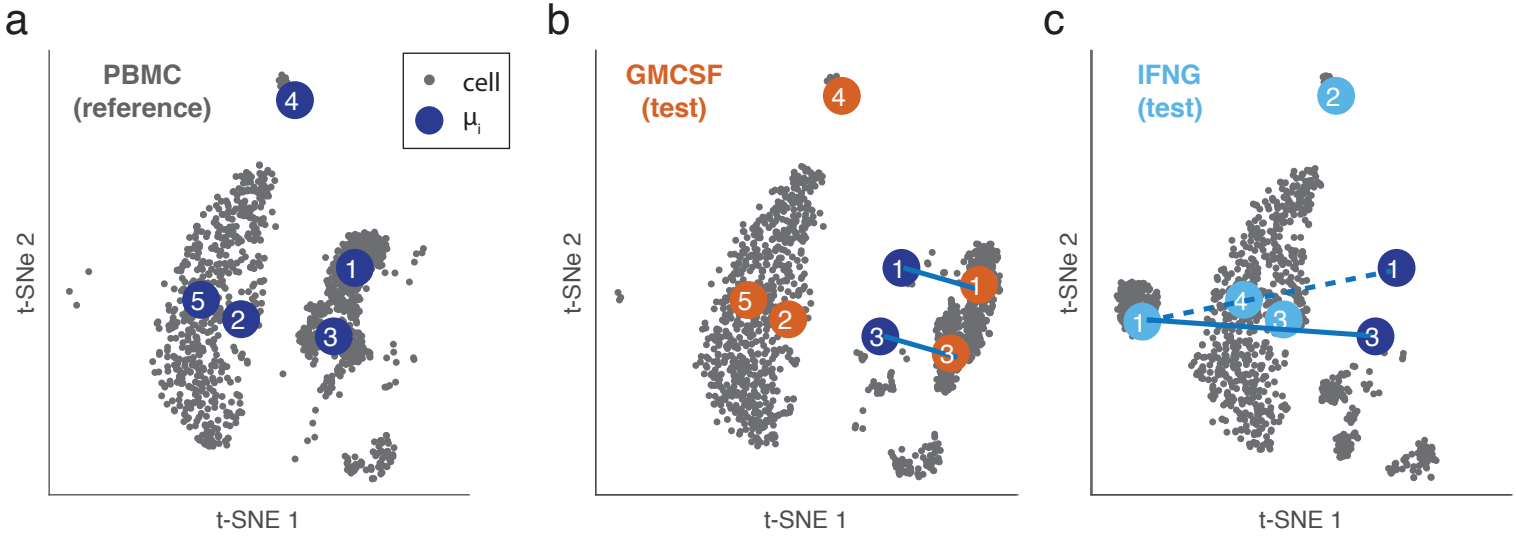

a
PBMC
(reference)
cell
μ_i
t-SNe 2
t-SNE 1
b
GMCSF
(test)
t-SNe 2
t-SNE 1
c
IFNG
(test)
t-SNe 2
t-SNE 1

Table 1: Patient Information

Patient Label	Age	Gender	Ethnicity	Disease	Medications
healthy1	58	Male	African American	none	N.D.
healthy2	43	Female	Caucasian	none	N.D.
MM1	51	Male	Caucasian	multiple myeloma	N.D.
MM2	43	Female	African American	multiple myeloma	N.D.
MM3	65	Male	Caucasian	multiple myeloma	Carflizomib/Zometa every 3 months Lisinopril Acyclovir
MM4	67	Male	Hispanic	multiple myeloma	Revlimid and Dexamethasone Hydrocodone Acetaminophen Vitamin D3 Diphenoxylate-atropine Calcium Alendronate Acyclovir Amlodipine Fenofibrate Warfarin